UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Raimundo de Castro Soares

Mineração de Dados para Entender os Fatores de Influência da Qualidade Educacional do Maranhão

São Luís
2022

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO MARANHÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Mineração de Dados para Entender os Fatores de Influência da Qualidade Educacional do Maranhão

Raimundo de Castro Soares

São Luís, 2022

Raimundo de Castro Soares

Mineração de Dados para Entender os Fatores de Influência da Qualidade Educacional do Maranhão

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - da Universidade Federal do Maranhão, como requisito para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Ariel Soares Teles
UFMA/IFMA
Coorientador: Prof. Dr. Luciano Reis Coutinho
UFMA

São Luís
2022

Raimundo de Castro Soares

Mineração de Dados para Entender os Fatores de Influência da Qualidade Educacional do Maranhão/ Raimundo de Castro Soares. – São Luís, 2022.

81 f.

Orientador: Prof. Dr. Ariel Soares Teles

Dissertação (Mestrado) – Universidade Federal do Maranhão – UFMA
Programa de Pós-Graduação em Ciência da Computação, 2022.

1. Mineração de dados. 2. Educação. 3. Maranhão. I. Soares Teles, Ariel, orient. II. Título.

CDU XXX

Raimundo de Castro Soares

Mineração de Dados para Entender os Fatores de Influência da Qualidade Educacional do Maranhão

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Raimundo de Castro Soares e aprovada pela comissão examinadora.

Aprovada em 23 de junho de 2022.

BANCA EXAMINADORA

_____
Prof. Dr. Ariel Soares Teles (Orientador)
UFMA/IFMA

_____
Prof. Dr. Luciano Reis Coutinho (Coorientador)
UFMA

_____
Prof. Dr. Davi Viana dos Santos (Examinador Interno)
UFMA

_____
Prof. Dr. Rafael Dias Araújo (Examinador Externo)
FACOM/UFU

Dedico a Deus por me abençoar e me dar forças para conseguir meus objetivos. A minha esposa, por me apoiar sempre e estar ao meu lado nas minhas conquistas. Ao meu filho autista, Diego Castro.

Agradecimentos

Agradeço primeiramente a Deus, por ter me abençoado a chegar até aqui;

Á minha família: esposa, filhos e pais, pelo incentivo e apoio ao longo da vida para atingir minhas metas;

Aos amigos: todos aqueles que, em algum momento ou em conversas rápidas incentivaram e acreditaram no meu potencial e nos meus objetivos.

"Tudo tem o seu tempo determinado, e há tempo
para todo o propósito debaixo do céu.
(Bíblia Sagrada, Eclesiastes 3, 1)

Resumo

O estado do Maranhão apresenta índices baixos na qualidade da educação básica, conforme pode ser verificado através das avaliações de desempenho nacionais ao longo dos anos. O problema da baixa qualidade educacional pode ser abordado e investigado do ponto de vista de diversas áreas. Uma delas é a utilização de mineração de dados educacionais, que está cada vez mais presente em estudos científicos, e também é utilizada para dar suporte à tomada de decisão por elaboradores de políticas públicas. No entanto, as pesquisas com os dados de uma área ou determinada localidade ainda podem ser escassas, sendo o caso do estado do Maranhão. Esta pesquisa de mestrado objetiva entender quais são os fatores que influenciam na qualidade da educação pública do estado do Maranhão com base nos dados históricos. Para isso, foram utilizadas técnicas de mineração de dados, tais como análise exploratória de dados, análise de correlação, análise de fatores, regressão e árvore de decisão. Os dados utilizados são das escolas públicas de ensino médio do Maranhão. Os resultados desse estudo mostram um diagnóstico da situação educacional do estado, com fatores que influenciam significativamente e outros sem tanta importância no desempenho da educação.

Palavras-chave: Mineração de Dados Educacionais, Maranhão, Educação, Análise de Correlação, Análise de Fatores, Regressão, Árvore de Decisão.

Abstract

The state of Maranhão has low levels of quality in basic education, as can be seen through national performance assessments over the years. The problem of low educational quality can be approached and investigated from the viewpoint of several areas. One of them is the use of educational data mining, which is increasingly present in scientific studies, and is also used to support decision-making by public policy makers. However, research with data from an area or a particular location may still be scarce, as is the case in the state of Maranhão. This master's research aims to understand what are the factors that influence the quality of public education in the state of Maranhão. For this purpose, data mining techniques were used, such as exploratory data analysis, correlation analysis, factor analysis, regression and decision tree. The data used are from public high schools in Maranhão. The results of this study show a diagnosis of the state's educational situation, with factors that significantly influence and others that are not so important in the performance of education.

Keywords: Educational Data Mining, Maranhão, Education, Correlation Analysis, Factor Analysis, Regression, Decision Tree.

Lista de ilustrações

Lista de tabelas

Lista de Siglas

CBIE Congresso Brasileiro de Informática na Educação.

CRISP-DM Cross Industry Standard Process for Data Mining.

DF Distrito Federal.

ENADE Exame Nacional de Desempenho de Estudantes da Educação Superior.

ENCCEJA Exame Nacional para Certificação da Educação de Jovens e Adultos.

ENEM Exame Nacional do Ensino Médio.

IBGE Instituto Brasileiro de Geografia e Estatística.

IDEB Índice de Desenvolvimento da Educação Básica.

IDH Índice de Desenvolvimento Humano.

IDH-M Índice de Desenvolvimento Humano Municipal.

IEMAs Institutos Estaduais de Educação, Ciência e Tecnologia do Maranhão.

INEP Instituto Nacional de Estudos Pedagógicos Anísio Teixeira.

IVS Índice de Vulnerabilidade Social.

KDD Knowledge Discovery in Databases.

LDIF Linked Data Integration Framework.

MDE Mineração de Dados Educacionais.

MEC Ministério da Educação.

PDDE Programa Dinheiro Direto na Escola.

PIB Produto Interno Bruto.

PPGCC Programa de Pós-graduação em Ciência da Computação.

RBIE Revista Brasileira de Informática na Educação.

RENOTE Revista Novas Tecnologias na Educação.

RMSE Erro Médio Quadrático.

RSL Revisão Sistemática de Literatura.

SAEB Sistema de Avaliação da Educação Básica.

SBIE Simpósio Brasileiro de Informática na Educação.

Sumário

1 Introdução

1.1 Contexto Geral

A mineração de dados, através da junção da estatística e a inteligência computacional (NAMEN; BORGES; SADALA, 2013), e usando metodologias próprias, trata e processa bases de dados, visando extrair informações relevantes (TAN; STEINBACH; KUMAR, 2009) e encontrar padrões relacionados a eles (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Ela está cada vez mais sendo utilizada nas diversas áreas do conhecimento, tais como educação, saúde e economia. Considerando a mineração de dados como uma área que realiza o processamento de bases de dados com técnicas apropriadas, pode-se afirmar que todas as áreas que dispõem de dados suficientes podem fazer uso dela.

A mineração de dados dispõe de ferramentas que podem analisar os mais diversos conjuntos de dados em busca de padrões e evidências, e proporcionar a descoberta de conhecimento não disponíveis em ferramentas comuns de banco de dados (AGARWAL, 2013). Quando essa análise é feita com dados da educação, buscando padrões em estudantes, predizendo resultados, identificando fatores, formas de aprendizagem e de interação, é chamada de Mineração de Dados Educacionais (MDE) (PEñA-AYALA, 2014).

O Ministério da Educação (MEC), através de avaliações e questionários direcionados a professores, alunos e gestão escolar, produz diversos dados relevantes (FONSECA; NAMEN, 2016), tais como quantidade de alunos matriculados, dados socioeconômicos, dependência administrativa das escolas, formação dos professores, localização das escolas, dentre outros. Esses dados são fontes de informações que, quando analisados detalhadamente, podem evidenciar o diagnóstico educacional em diversos aspectos. O Instituto Nacional de Estudos Pedagógicos Anísio Teixeira (INEP) é uma agência do MEC que disponibiliza publicamente bases de dados sobre a educação nacional. Os dados disponibilizados pelo INEP, quando analisados, podem mostrar a realidade boa ou ruim da educação brasileira. Essa análise pode ser feita através de técnicas e algoritmos de mineração de dados.

O diagnóstico da educação pode ser feito a partir de uma série de indicadores, como o desempenho escolar, a distorção entre idade e série, bem como a progressão dos alunos ao longo dos anos (i.e., a eficiência do sistema de ensino por meio do fluxo escolar), as condições oferecidas pela rede de ensino, os recursos disponíveis na escola, e a qualificação dos professores (RIGOTTI; CERQUEIRA, 2015). No Brasil, o IDEB é o indicador utilizado para medir a qualidade da educação básica.

A pontuação média do IDEB no Maranhão nos últimos anos, referente às escolas

públicas de ensino médio, melhorou, mas não significativamente. O último índice, referente ao ano de 2019 (3,7), ainda está abaixo da nota dos estados com melhor pontuação (>=4,0), ou mesmo da nota média nacional (3,9). O motivo pelo qual o IDEB no Maranhão não evoluiu significativamente é algo que deve ser estudado, buscando identificar os possíveis fatores que mais influenciam o desempenho das escolas do estado.

Uma análise dos dados relacionados ao IDEB pode gerar evidências que expliquem o problema do baixo rendimento escolar, ao menos em parte, e contribuir para a tomada de decisão dos gestores educacionais. Neste contexto, esta pesquisa de mestrado aborda o estudo dos fatores que influenciam a qualidade da educação maranhense, utilizando mineração de dados educacionais.

1.2 Caracterização do Problema

Há uma carência, em particular na literatura científica, em realizar o processo de mineração em bases de dados importantes, como os dados educacionais do Estado do Maranhão. Uma hipótese para a qual essa análise de dados não foi realizada é a falta de conhecimento sobre a disponibilidade desses dados. Também, a falta de recursos humanos é uma possibilidade para que a análise dos dados não tenha ocorrido ainda. Nesse sentido, a ausência de estudos voltados à mineração de dados educacionais no Maranhão demonstra uma subutilização dos dados disponíveis, evidenciando a necessidade de pesquisas.

O estado do Maranhão apresenta baixos índices educacionais, caracterizando uma educação com desempenho ruim. Esses baixos indicadores podem ter várias causas, que precisam ser investigadas e estudadas. Em particular, o IDEB do Maranhão precisa ser melhorado, sendo necessário ser estudado para a identificação de eventuais fatores que podem proporcionar sua melhoria. Dessa forma, o problema de estudo desta pesquisa mestrado está relacionado ao fato de não haver um conhecimento profundo, sob o aspecto da mineração de dados, dos fatores que causam impactos e influência na qualidade educacional do estado do Maranhão. O problema da baixa qualidade educacional maranhense pode ser causado por diversos fatores e esta pesquisa procura identificar e explicar eles.

1.3 Hipótese de Pesquisa

A hipótese deste trabalho de mestrado é que, ao tirar vantagem de diferentes técnicas de mineração de dados educacionais (e.g., análise de correlação e de fatores, modelos de regressão e árvore de decisão) com as bases de dados do INEP, especificamente o conjunto de dados do SAEB e do IDEB, é possível identificar os fatores que influenciam na qualidade educacional do Maranhão.

1.4 Relevância do Trabalho

A educação é um dos pilares que implica uma melhor condição de vida da sociedade em geral. Por meio da educação, é possível melhorar o índice de desenvolvimento humano, reduzir o desemprego, proporcionar melhores salários e melhores condições de trabalho. No entanto, para cumprir esse propósito, a educação precisa ter qualidade.

A educação pode ser influenciada por diversos fatores, que podem causar impactos positivos ou negativos. Por exemplo, evasão e reprovação de alunos são itens que impactam negativamente na educação. Além disso, esses fatores podem contribuir para a diminuição dos índices de qualidade educacional. Já a participação dos pais e da família, dando incentivo aos alunos, são itens que podem impactar positivamente na educação.

Uma vez identificados os fatores que podem influenciar na qualidade educacional, pode-se criar políticas públicas para alavancar esses fatores no sentido de fazer com que eles sejam melhorados ou aperfeiçoados para causar um impacto ainda melhor e mais evidente da educação na sociedade. Ou, caso o fator seja negativo (e.g., evasão, reprovação), pode ser estudado sua origem, para ser combatido.

Embora existam diversos trabalhos que abordam a qualidade educacional (JÚNIOR *et al.*, 2019; PENTEADO, 2016b; SOARES, 2006; SOARES *et al.*, 2021), buscando identificar as causas para os índices baixos, esta pesquisa de dissertação se faz necessária, pois, para o melhor do nosso conhecimento, não foram identificadas pesquisas semelhantes, utilizando mineração de dados sobre a educação do estado do Maranhão.

1.5 Objetivos

O objetivo geral desta pesquisa de mestrado é identificar os fatores de influência na qualidade educacional do Maranhão através do uso de técnicas de mineração das bases de dados fornecidas pelo INEP.

Para tanto, consideram-se os seguintes objetivos específicos:

- Realizar uma RSL para levantar o estado da arte referente à mineração de dados com as bases de dados educacionais do IDEB;
- Realizar a análise exploratória dos dados do ensino médio da educação pública do Maranhão;
- Identificar os fatores de influência na nota do IDEB do ensino médio das escolas públicas estaduais do Maranhão;
- Criar modelos exploratórios para estudar e explicar o ensino médio da educação pública estadual maranhense.

1.6 Metodologia de Pesquisa

As metodologias empregadas em mineração de dados são utilizadas para uma melhor otimização do processo, para seguir uma sequência de etapas e procedimentos e, dessa forma, conseguir atingir os objetivos pretendidos. A metodologia CRISP-DM foi adotada para realizar esta pesquisa. Ela é possível de ser adaptada a qualquer categoria de negócio e trata-se de uma metodologia interativa, ou seja, executam-se suas etapas, podendo voltar para uma etapa anterior caso esta precise ser refeita. Ela está dividida em 6 etapas, ilustradas na Figura 1 e explicadas em seguida (AZEVEDO; SANTOS, 2008; SHEARER, 2000).

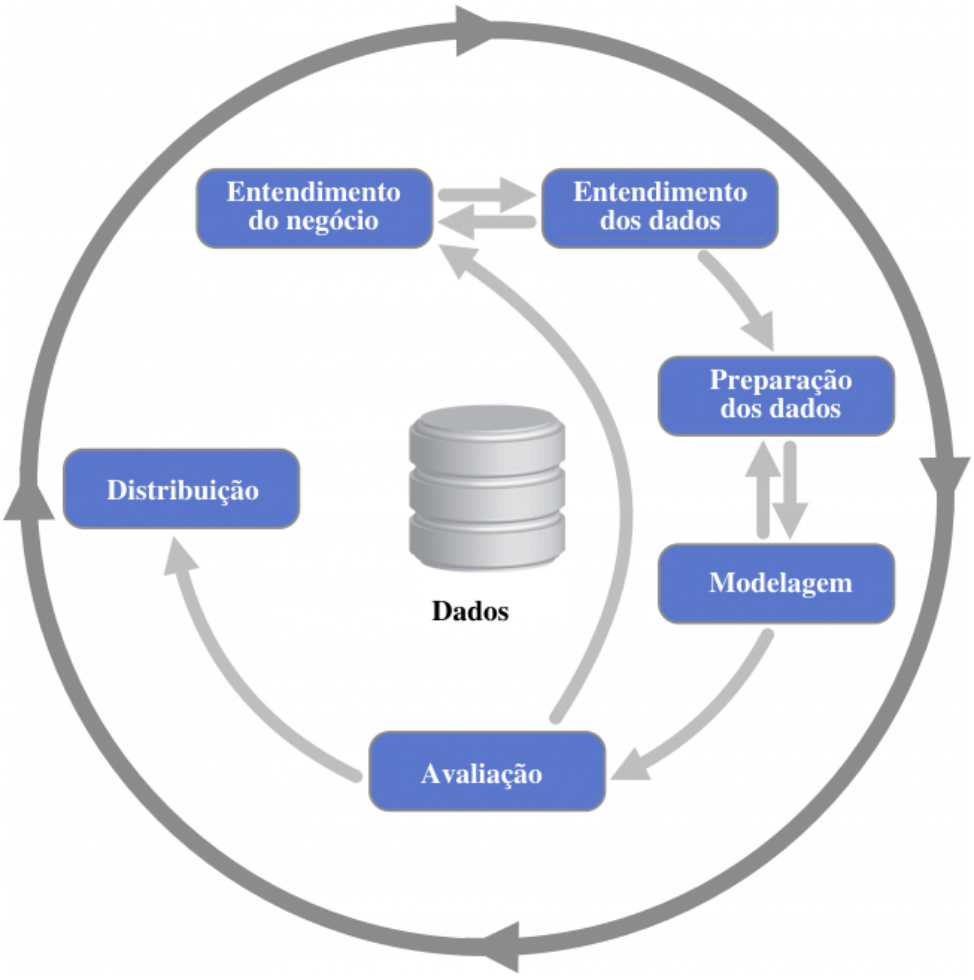

Figura 1 – Ciclo da Metodologia CRISP-DM, adaptado de (SHEARER, 2000).

- **Entendimento do negócio:** nesta primeira etapa é feito o entendimento dos objetivos do negócio, bem como a elaboração de perguntas na perspectiva de encontrar a solução através da mineração de dados;
- **Entendimento dos dados:** nessa etapa é feita a coleta dos dados e já se inicia uma análise prévia para seu entendimento. Eventualmente, poderão ser identificados problemas e os primeiros *insights*;

- **Preparação dos dados:** esta etapa envolve a preparação da base de dados que podem ter que passar por diversos processos, até que os dados estejam prontos para serem minerados. A preparação pode incluir: padronização e normalização dos dados, exclusão de variáveis, junção com outras bases, alteração de nomes de variáveis para um melhor entendimento, e eventuais outras modificações necessárias na base de dados;

- **Modelagem:** nesta fase ocorre a elaboração do modelo para a análise dos dados (e.g., árvores de decisão, regressão). Diversos modelos podem ser criados, analisados e testados, e posteriormente escolher um deles (ou mais de um) para resolver o problema;

- **Avaliação:** nesta etapa é feita uma verificação se o modelo (ou os modelos, em caso de mais de um) adotado respondeu às questões do problema elaboradas na primeira fase. Caso a avaliação não seja satisfatória, nessa etapa é possível voltar para fase inicial para refazer o processo, ou parte dele. Isso explica a iteração presente na metodologia CRISP-DM;

- **Distribuição:** entrega dos resultados da análise dos dados, feita através de relatórios, imagens, gráficos ou ilustrações.

1.7 Organização do Trabalho

O restante desta dissertação de mestrado está organizado como descrito a seguir. A Figura 2 apresenta a organização dos capítulos da dissertação, dando destaque para os capítulos que representam as 5 etapas executadas da metodologia CRISP-DM. A sexta etapa, que é a distribuição, em que são apresentados os resultados da mineração, é realizada através desta dissertação e dos artigos publicados originados da pesquisa. Convém ressaltar também que esta dissertação não possui um capítulo específico de Fundamentação Teórica, pois, os conceitos teóricos utilizados neste trabalho estão explicados ao longo de todos os capítulos.

- O **capítulo 2** discute os trabalhos relacionados que se assemelham, ou estudam tópicos similares, a este estudo;

- O **capítulo 3** expõe o entendimento do negócio e entendimento dos dados, sendo a primeira e segunda etapa da metodologia, respectivamente;

- O **capítulo 4** explica a preparação e modelagem dos dados, sendo a terceira e quarta etapa da metodologia.

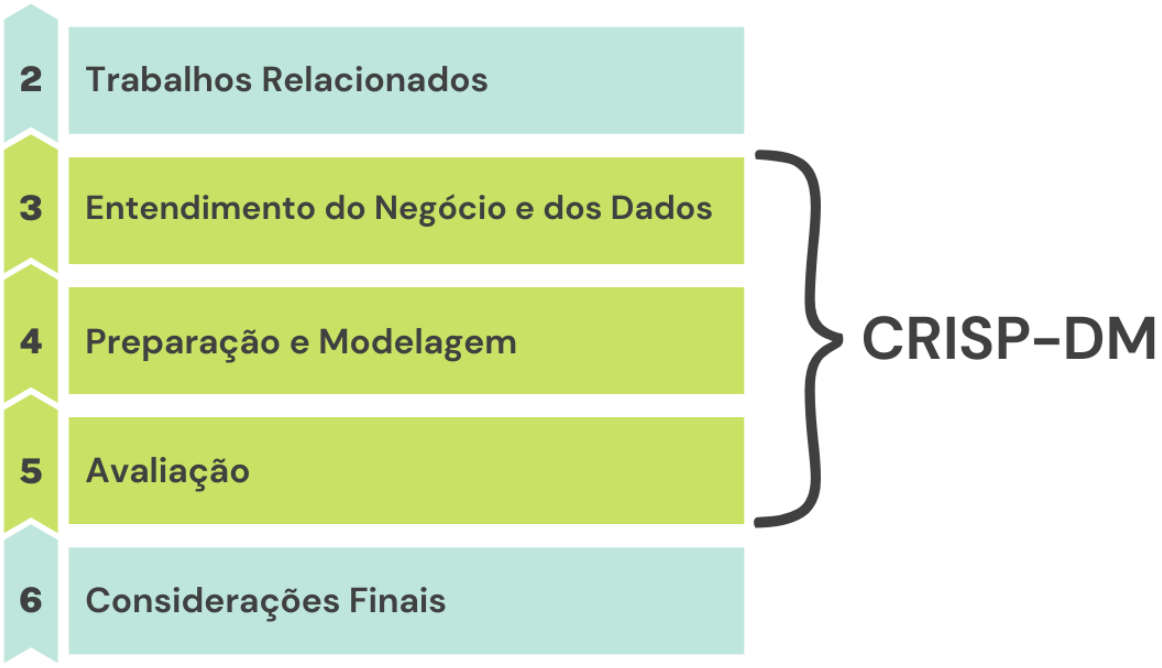

Figura 2 – Organização da dissertação seguindo as etapas da metodologia CRISP-DM.

- O **capítulo 5** apresenta a avaliação, correspondendo à quinta etapa da metodologia, com as respostas das questões de pesquisa, os principais achados, dificuldades encontradas na pesquisa, limitações do estudo e trabalhos futuros;
- O **capítulo 6** finaliza a dissertação com as considerações finais, as principais contribuições desta pesquisa e a listagem dos artigos publicados/submetidos ou em produção.

2 Trabalhos Relacionados

Os trabalhos relacionados a esta pesquisa foram identificados e analisados através de uma RSL, onde foi realizado o levantamento do estado da arte referente à mineração de dados com as bases disponibilizadas pelo INEP. O artigo com os resultados detalhados desta RSL (SOARES *et al.*, 2021) foi publicado na Revista Novas Tecnologias na Educação (RENOTE) (Apêndice A, página 71). O processo sistemático e os resultados desta revisão são apresentados de forma sumarizada a seguir.

2.1 Metodologia da RSL

A busca foi realizada nas seguintes bibliotecas digitais, bases de artigos e anais de eventos: ACM Digital Library, IEEE Xplore, Science Direct, Web of Science, Workshops do Congresso Brasileiro de Informática na Educação (CBIE), Revista Brasileira de Informática na Educação (RBIE), RENOTE e Simpósio Brasileiro de Informática na Educação (SBIE). Para selecionar os estudos primários, os seguintes critérios de inclusão foram aplicados: (i) artigos que realizam mineração dos dados do INEP (i.e., SAEB, IDEB, Censo Escolar ou Indicadores Educacionais); (ii) artigos em inglês ou português; e (iii) trabalhos completos. Os critérios de exclusão foram: (i) artigos que realizam mineração de dados que não sejam do SAEB, IDEB, Censo Escolar e Indicadores Educacionais; (ii) artigos em idioma diferente do inglês e português; (iii) literatura cinza (e.g., teses, dissertações, artigos curtos, capítulos de livro, relatórios técnicos); e (iv): texto completo não disponível online.

Para busca dos artigos foram utilizadas as seguintes strings:

- A string (“data analytics” OR “data mining” OR “data analysis”) AND (SAEB OR INEP OR IDEB OR “basic education” OR “school census” OR “educational indicators”) AND (Brazil OR Brazilian) foi utilizada nas fontes: ACM Digital LIbrary, IEEE Digital Library, Science Direct e Web of Science.
- A string (dados OR data) AND (SAEB OR INEP OR IDEB OR “educação básica” OR “basic education” OR “censo escolar” OR “indicadores educacionais”) foi utilizada nas fontes: CBIE, RBIE, RENOTE e SBIE.

A pesquisa nas bases de dados resultou em um total de 410 publicações, com apenas 3 artigos duplicados e removidos. Houve então a aplicação dos critérios de seleção aos 407 artigos. Após a seleção realizada, um total de 19 artigos foram selecionados. Com base nessa RSL e nas informações obtidas a partir da extração de dados dos artigos, foi possível identificar os trabalhos relacionados. A seguir são apresentados os trabalhos relacionados

a este estudo, que fazem mineração de dados educacionais nacionais e regionais (i.e., de um estado ou de uma cidade).

2.2 Descrição dos Trabalhos Relacionados

O autor Soares (SOARES, 2006) desenvolveu uma pesquisa utilizando a base de dados do SAEB para estudar o baixo desempenho escolar constatado através das desigualdades educacionais de desempenho cognitivo na disciplina de Matemática da educação básica nacional. Para medição do desempeno cognitivo utilizado no estudo, o pesquisador usou ideias econômicas de concentração de renda, medida pelo coeficiente de Gini, e medidas de pobreza. A pesquisa utilizou a metodologia de análise exploratória com técnica de análise estatística. Conforme os dados utilizados, para fins de estudos os alunos foram agrupados por raça, sexo, nível socioeconômico (SES) e região de residência.

Constatou-se uma grande diferença no nivel de proficiência entre alunos do quarto quartil (altos valores) e os alunos do primeiro quartil (baixos valores). O pesquisador também verificou que as desigualdades educacionais do Brasil são mais evidentes que as desigualdades econômicas. Para comparar o desempenho de alunos de escolas públicas com os da privada, o pesquisador verificou que as proficiências dos alunos de escolas particulares são imensamente maiores. Mas isso não se deve exatamente a poder econômico, e sim a aspectos como estilo de gerenciamento da escola, projeto pedagógico, estilo do professor e melhor envolvimento do setor privado na educação. No que diz respeito aos grupos separados por sexo, no geral, os alunos do gênero masculino são melhores que as meninas em matemática. Quanto aos grupos por regiões, as regiões norte e nordeste possuem os piores desempenho, enquanto os melhores desempenhos são de alunos das regiões sul, sudeste e centro-oeste.

Os pesquisadores Waltenberg e Vandenberghe (WALTENBERG; VANDENBERGHE, 2007) utilizaram a base de dados do SAEB para estudar o baixo desempenho escolar. A pesquisa focou em estudar os gastos por aluno e seus componentes que consomem esses gastos: salário do professor e tamanho da turma (número de alunos na sala). Através da metodologia de análise exploratória, foram aplicadas as técnicas de Algoritmo EOp e Modelo de Regressão. De acordo com as descobertas dos pesquisadores, foram analisadas as relações de gastos necessários para promover uma equalização de oportunidades.

Os pesquisadores descobriram que, para que seja promovida e implantada uma política igualitária entre alunos de diferentes origens socioeconômicas, partindo da premissa dos valores investidos por alunos, é necessário multiplicar esse valor por 6,8 em média, considerando o valor de gastos no momento da pesquisa (ano de 2007). Os autores concluíram que os gastos por alunos aumentam bem pouco ao longo dos anos e o investimento diretamente em dinheiro e em outros aspectos, como infraestrutura, por exemplo, pode

reduzir consideravelmente uma possível redistribuição financeira.

Os autores Ramos et al. (RAMOS; MACHADO; CORDEIRO, 2015) utilizaram as bases de dados do SAEB e IDEB para estudar o problema da fragilidade do sistema educacional. Através de análise exploratória, foi aplicada a técnica de análise de agrupamento. A justificativa para está técnica é que devido ao tamanho da base de dados, e sendo o Brasil um país continental, os dados foram separados em grupos para facilitar a análise e não permitir conclusões generalizadas. Após a construção do banco de dados agrupados, iniciou-se a análise para entender como escolas do país são e entender possíveis discrepâncias apresentadas nas notas do IDEB e características dos grupos formados. Os grupos foram separados por características semelhantes. O primeiro grupo foi criado com as escolas distribuídas por anos, formando quatro grupos referentes aos anos de 2007, 2009, 2011 e 2013. E os resultados mostraram que houve uma distribuição muito semelhante das escolas para todos os quatro anos. No grupo separado por estados da federação, percebe-se que as escolas têm distribuição aproximada à população. No grupo por tipos de escolas (i.e., municipal, estadual, federal), a maioria da amostra é composta por escolas municipais, pois essa esfera é responsável por grande parte das escolas do país. Em relação às escolas federais, elas têm desempenho superior por serem formadas pela classe alta do ensino público do país, em função do pequeno número de escolas e maior capacidade de captação de recursos.

Os autores também agruparam todas as escolas e então identificaram possíveis melhores e piores desempenhos por estado e outras características socioeconômicas das regiões para explicar o desempenho. Os estados de São Paulo e Minas Gerais possuíram o maior investimento em educação, com um maior salário para os professores desses estados. Além disso, o estado de Minas Gerais também teve o melhor desempenho devido ao maior investimento no estado e ao menor número de alunos por professor (24,8 em Minas Gerais versus 30,4 em São Paulo).

Penteado (PENTEADO, 2016a) utilizou os dados nacionais do SAEB de 2013, referente à disciplina de matemática do 9º ano do ensino fundamental para extrair relações de pré-requisitos entre habilidades avaliadas em avaliações de larga escala. Ou seja, se uma habilidade A depende da habilidade B, se uma habilidade é pré-requisito para outra. O experimento foi realizado usando R com o algoritmo Partial Order Knowledge Structure (POKS).

O algoritmo considera cada par possível de habilidades, verificando se existe interação entre elas, por meio das respostas de desempenho medidas pelos itens correspondentes a cada habilidade (acertos e erros). De acordo com o pesquisador, para se assumir que A é pré-requisito de B, supõe-se que: se o aluno não dominar a habilidade A, muito provavelmente não dominará B também. De modo análogo, se dominar B, muito provavelmente dominará A. Como resultado da pesquisa, algumas habilidades são mais básicas, com pouca

dependência de outras habilidades. No entanto, se destacaram duas habilidades que mais dependem uma da outra, ou que outras dependem delas: habilidade 36 (Resolver problema envolvendo informações apresentadas em tabelas e/ou gráficos) e habilidade 37 (Associar informações apresentadas em listas e/ou tabelas simples aos gráficos que as representam e vice-versa). Também foram encontradas habilidades isoladas, sem dependência uma das outras. Nesse quesito, os destaques são para as habilidades 06 (Reconhecer ângulos como mudança de direção ou giros, identificando ângulos retos e não-retos) e habilidade 13 (Resolver problema envolvendo o cálculo de área de figuras planas).

Outro trabalho realizado pelo pesquisador Penteado (PENTEADO, 2016b) utilizou a base de dados do SAEB para estudar o problema do baixo desempenho escolar. Ele realizou uma análise de correlação entre o desempenho escolar (dados do IDEB) e os indicadores municipais. Os indicadores socioeconômicos (IDH-M, Índice de Vulnerabilidade Social (IVS), índice de Gini) foram extraídos do Instituto de Pesquisa Econômica Aplicada (IPEA). O pesquisador utilizou as variáveis de desempenho acadêmico como variáveis dependentes, e as variáveis socioeconômicas como variáveis independentes para análise de suas correlações.

Após os dados preparados, foi calculada a matriz de correlação relacionando os indicadores e os desempenhos do IDEB, utilizando a correlação de Pearson entre as variáveis. Como resultado, notou-se forte correlação (magnitude variando de 0,4 a 0,7) do IVS e IDH-M com o IDEB para 5º e 9º anos, tanto para escolas municipais quanto estaduais, tendo maior influência no ensino fundamental (5º ano, tanto municipal quanto estadual). O coeficiente de Gini apresentou correlação moderada com o IDEB. O Produto Interno Bruto (PIB) per capita também apresentou correlação moderada para o IDEB municipal e um pouco menor para as escolas estaduais.

Os autores Júnior et al. (JÚNIOR *et al.*, 2017) utilizaram a base de dados do censo escolar e do Exame Nacional do Ensino Médio (ENEM) para realizar a mineração e identificação de correlação e *outliers*, especificamente de dados do estado de Pernambuco. Essa pesquisa visou analisar a correlação entre características das escolas e o desempenhos dos alunos no ENEM. A correlação indica o relacionamento linear, ou não linear, entre duas variáveis. Portanto, em um conjunto de dados com diversas variáveis, podem ser medidas várias correlações. Os índices de correlação entre características das escolas e as notas dos alunos no ENEM podem se apresentar de forma baixa, moderada e alta. Essa classificação (COHEN, 2013) considera escores de correlação entre 0,10 e 0,29 como pequenos (baixos), escores entre 0,30 e 0,49 podem ser considerados médios (moderados) e escores entre 0,50 e 1,0 podem ser considerados grandes (altos). Outliers são informações com valores atípicos e pode ser usada para identificar desigualdades educacionais por região, por exemplo. Para identificação de outliers, podem ser analisados os quartis superiores, onde se encontram 25 por cento dos valores mais elevados de determinada variável, ou os

quartis inferiores, onde se encontram 25 por cento dos menores valores da variável.

Ao realizar a correlação das notas do ENEM, para cada área do conhecimento, o índice de correlação foi superior a 0,7. Os pesquisadores concluíram que, seja qual for a área do conhecimento, se a nota de um aluno aumentar em uma determinada área, também tende a aumentar em outra. Com os dados do censo escolar, foram agrupadas as escolas com as características de: água filtrada, água da rede pública, sistema de esgoto e coleta de lixo. Ao fazer a correlação das escolas que possuíam ou não estas características, com as médias gerais dos alunos destas no ENEM, o resultado da correlação foi 0,276. Com essa correlação baixa, os pesquisadores concluíram que a presença ou não destas características em uma escola não influenciam a nota do ENEM.

O relacionamento das escolas com laboratório de informática com a média geral do ENEM obteve um resultado de 0,478. Enquanto a correlação das escolas com laboratório de informática com as notas específicas de ciências da Natureza e Matemática, obteve os respectivos valores de 0,453 e 0,432. Os autores concluíram que a correlação entre a presença de laboratório de informática e as melhores notas pode ser considerada moderada. A mesma análise foi realizada, fazendo a correlação das escolas com laboratório de ciências, com as notas gerais do ENEM, e das áreas de Ciências da Natureza e Matemática. Os valores obtidos foram 0,553, 0,504 e 0,507, respectivamente. Com esses valores, os autores concluíram haver uma alta correlação entre a presença do laboratório de ciências e a melhoria dessas três notas dos alunos no ENEM.

A análise entre a presença de biblioteca com as notas gerais e as notas de Linguagens, obteve os valores baixos de correlação, de 0,244 e 0,289, respectivamente. No entanto, a correlação entre a presença de biblioteca e a nota específica da redação, obteve a correlação moderada de 0,362. Ao realizar a análise entre a presença de salas de leitura e as médias gerais da prova do ENEM, nota de Redação e da área de Linguagens, foram obtidos valores altos de correlação de 0,615, 0,557 e 0,557, respectivamente. Os autores afirmaram que a presença de sala de leitora indica a possível melhora nas notas do ENEM nessas áreas comparadas. Ao tentar identificar outliers, não foram encontradas discrepâncias entre as notas do ENEM e as características de infraestrutura das escolas. As escolas e notas analisadas apresentaram características semelhantes.

Carvalho et al. (CARVALHO; CRUZ; GOUVEIA, 2017) utilizaram a base de dados dos Censos da Educação Básica e Superior, referentes aos anos de 2014 e 2015, para estudar o problema da evasão de alunos no âmbito do estado de Pernambuco. A mineração de dados foi realizada pela metodologia Knowledge Discovery in Databases (KDD) com o software WEKA, que dispõe de algoritmos de mineração de dados. Os algoritmos usados pelos pesquisadores foram: Árvore de Decisão (Algoritmo J48) e Classificação Bayesiana (Naive Bayes). Segundo os autores, a escolha do Algoritmo J48 considerou o motivo que a técnica de classificação por árvore de decisão gera regras objetivas que facilitam a interpretação

dos resultados (WITTEN *et al.*, 2005). Outro motivo para a escolha do J48 diz respeito às taxas de acurácia de boa aceitação. Já o algoritmo Naive Bayes foi escolhido por elaborar classificação probabilística simples calculando o conjunto de probabilidades contanto a frequência e a combinação de valores do conjunto de dados (DIMITOGLOU; ADAMS; JIM, 2012; GONZALEZ *et al.*, 2012). Conforme afirmam os autores, o algoritmo considera todos os atributos independentes, dado o valor da variável de classe. Trata-se de um modelo estatístico que permite determinar a probabilidade de hipóteses ocorrerem em um determinado conjunto de dados. Como resultado da pesquisa, eles conseguiram elaborar perfis das escolas quanto à infraestrutura, perfis de alunos da educação básica de acordo com a região da escola em que estudam, perfis de alunos da educação superior dos cursos da área de Tecnologia da Informação e Comunicação, e também fizeram a comparação entre os perfis desses alunos que residiam na região metropolitana da capital (Recife) e os alunos do interior do estado.

Os pesquisadores Souza et al. (BEM; PEREIRA; SOUZA, 2017) utilizaram os dados do IDEB para a criação de um *data mart* com o objetivo de realizar a análise comparativa das cidades da microrregião do Pajeú, no estado de Pernambuco. *Data mart* é uma derivação originária de um *Data Warehouse*. Já um *Data Warehouse* (DW) refere-se a grandes bancos de dados originários de sistemas transacionais com disponibilização de uma estrutura de dados dimensionais com o propósito de dar apoio à tomada de decisão de gestores, possibilitando o processamento analático por meio de ferramentas especificas (BARBIERI, 2001). A solução desenvolvida foi capaz de mostrar os dados graficamente através de tabelas, com números, gráficos e cores, possibilitando a análise, comparação e tomada de decisão. A conclusão dos pesquisadores foi que os dados precisam ser disponibilizados de maneira facilitada e entendível, o que pode ser feito através da ferramenta utilizada.

Os pesquisadores Avila et al. (AVILA *et al.*, 2018) utilizaram os dados do Programa Dinheiro Direto na Escola (PDDE) e dos Indicadores Educacionais para produzir um *Mashup* de Dados, através de técnicas de Linked Data (dados Ligados e de Web Semântica), com o objetivo de facilitar o entendimento e disponibilização de dados que, originalmente, são disponibilizados em formatos diferentes. Os dados foram integrados através do framework Linked Data Integration Framework (LDIF) seguindo o seguinte fluxo: i) Extração das fontes de dados; ii) Exportação das visões e transformação dos dados; iii) Resolução da identidade através de links; iv) Avaliação da qualidade e fusão dos dados; e v) Saída dos dados (BIZER *et al.*, 2012). Forma utilizados dados nacionais referentes aos anos de 2016 e 2017.

Como resultado, os pesquisadores conseguiram elaborar e publicar o *Linked Data Mashup* sobre dados de indicadores da educação brasileira e de execuções financeiras do PDDE. Para validar o *mashup*, foram realizadas consultas SPARQL para a visualização dos

dados para análise. Como exemplo, uma consulta pode mostrar o repasse fornecido pelo PDDE e as respectivas taxas de aprovação e abandono. Para esse exemplo, os pesquisadores concluíram que, as escolas que possuem uma menor quantidade de alunos, por conseguinte, apresentam maiores taxas de abandono, refletindo no repasse do PDDE. Tal situação pode vir eventualmente a impactar na decisão de pais e alunos durante a escolha de qual instituição de ensino o aluno ira efetuar sua matrícula.

Os autores Nascimento e Júnior (NASCIMENTO; JÚNIOR, 2018) utilizaram as bases de dados nacionais para, através de técnica de regressão, estimar o indicador de professores com cursos superiores que atuam na educação básica. A metodologia utilizada foi a CRISP-DM, e foram elaboradas dois tipos de regressão: regressão linear simples e regressão linear robusta. O erro da regressão robusta possuiu menor valor em comparação à regressão linear. Apesar de o desvio padrão do erro da regressão linear ser um pouco inferior, não consiste em uma diferença significativa.

Os autores em (NASCIMENTO; JUNIOR; FAGUNDES, 2018) utilizaram a base de dados dos Indicadores Educacionais para estudar o problema da evasão e reprovação de alunos. Eles utilizaram a metodologia CRISP-DM. Após realizarem a análise de correlação das variáveis, elaboraram modelos de regressão linear e robusta, e compararam o desempenho dos dois tipos de regressão para verificar qual deles melhor minimiza o erro de previsão. A qualidade dos modelos de regressão foi avaliada pelo erro médio absoluto e desvio padrão. Na predição dos indicadores de evasão e reprovação, a regressão robusta obteve o melhor desempenho em comparação com a regressão linear.

Os pesquisadores Nascimento et al. (NASCIMENTO; FAGUNDES; MACIEL, 2019) realizaram um estudo de previsão utilizando a base dos indicadores educacionais para estudar o problema do desempenho previsto nas taxas de eficiência. Eles utilizaram a metodologia CRISP-DM e modelo de regressão. As taxas de eficiência escolar considderam 3 itens: evasão, reprovação e aprovação.

Os pesquisadores utilizaram dois modelos de regressão: regressão liunear multipla e regressão linear robusta (paramétrica e não paramétrica. Como resultados, o primeiro modelo de regressão minimizou o erro deprevisão para os indicadores educacionais de evasão, reprovação e aprovação na maioria aobsoluta dos csos de escolas de ensino fundamental e ensino médio. Os pesquisadores concluíram que o uso de regressão em previsão de indicadores educacionais são capazes de elaborar modelos eficientes de estimativa de variáveis educacionais. Além disso, essas ferramentas podem ser usadas amplamente na geração de conhecimento e auyxiliando na solução de problemas e criação de mecanismos que auxiliam e apoiam o ensino e aprendizagem.

Os pesquisadores Pinto et al. (PINTO; JÚNIOR; COSTA, 2019) utilizaram a base de dados do SAEB para identificar os fatores que afetam o desempenho escolar dos alunos (notas do IDEB) dos anos finais (9º ano) do ensino fundamental das escolas

públicas municipais da cidade de Teotônio Vilela, estado de Alagoas, referente aos anos de 2015 e 2017. A metodologia CRISP-DM foi utilizada pelos autores para a realização do estudo. Visando gerar dados sintéticos para equilibrar a base de dados para as variáveis dependentes, foi utilizada técnicas de balanceamento de dados através do *Synthetic Minority Oversampling Techniques* (SMOTE). O método consiste na geração de mais dados das classes minoritária por meio da adição de instâncias próximas.

A pesquisa identificou 18 atributos que mais influenciam o desempenho dos alunos nas disciplinas de Português e Matemática. No que diz respeito aos algoritmos capazes de identificar a melhor precisão de classificação dos atributos, os algoritmos OneR, LibSVM e J48 apresentaram acurácia acima de 98%, constituindo, portanto, os melhores resultados em classificação. O estudo identificou os atributos que mais influenciam nas notas do IDEB dos alunos da rede pública municipal de Alagoas e, dessa forma, os autores concluíram que, uma vez conhecendo esses atributos, pode-se ter uma ideia de como melhorar os índices educacionais.

Os pesquisadores Junior et al. (JUNIOR *et al.*, 2019) usaram as bases de dados do INEP para elaborar um modelo de predição dos índices de aprovação e reprovação de alunos no ensino médio do estado de Pernambuco. O modelo para a estimação dos indicadores foi realizado por meio da Regressão Quantílica Não Paramétrica (RQNP) e Otimizada por Algoritmos Genéticos (AG), através do software R. O resultado foi comparado entre as estimações realizadas pelo modelo RQNP padrão e RQNP otimizado por Algoritmos Genéticos. De acordo com os resultados de mediana e desvio padrão, percebeu-se que o modelo otimizado por AG obteve uma mediana de erro menor com maior desvio padrão. Para avaliar a confiabilidade estatística dos modelos, os autores realizaram o teste estatístico de Wilcoxon não pareado com 5% de significância. O resultado alcançado pelo teste indicou que os resultados alcançados pelos modelos são diferentes. Os modelos propostos neste trabalho podem ser utilizados como sistemas de apoio a decisão, já que indicam em que casos pode haver uma reprovação.

Freitas Júnior et al. (JÚNIOR *et al.*, 2019) realizaram um trabalho de mineração de dados com as bases de dados do IDEB e do SAEB para estudar o problema da baixa qualidade da gestão escolar e baixo desempenho escolar das escolas municipais de Maceió, no estado de Alagoas. A metodologia utilizada foi a CRISP-DM, e usando modelos de regressão e árvore de decisão. Os dados utilizados foram da rede pública municipal referente aos anos iniciais do ensino fundamental (1º ao 5º) ano. Os pesquisadores realizaram os experimentos através do software WEKA. Para realização da pesquisa, foram utilizados dois algoritmos: Regressão Linear Simples e o J48 para árvore de decisão. A partir da árvore de decisão elaborada, os autores concluíram que a infraestrutura da escola não tem tanta influência no IDEB, enquanto as ações pedagógicas dos professores em sala de aula e da gestão escolar estão muito mais ligadas à melhoria das notas do IDEB.

Outro trabalho relacionado foi realizado pelos pesquisadores Pinto et al. (PINTO *et al.*, 2019), também semelhante a (JÚNIOR *et al.*, 2019) (trabalho descrito anteriormente). Através de mineração de dados eles analisaram os resultados de avaliações oficiais realizadas pelo (INEP para analisar a influência no desempenho do (IDEB. O estudo recorreu à metodologia (CRISP-DM e trabalhou com dados de 13 escolas da rede pública municipal do ensino médio da capital do estado de Alagoas (Maceió). A pesquisa identificou 10 atributos que mais influenciam o desempenho dos alunos nas disciplinas Português e Matemática. Os algoritmos utilizados J48, OneR, JRip e LibSVM tiveram resultados com 100 por cento de acurácia de classificação para os dados das provas de português e matemática.

A pesquisadora Pacini (PACINI, 2020) realizou uma pesquisa de mineração de dados para analisar os indicadores educacionais de esforço, regularidade e adequação da formação do docente em relação à média de proficiência da Prova Brasil de Língua Portuguesa e Matemática, na edição do ano de 2015, do 5º e 9º ano do Ensino Fundamental, da rede estadual de ensino do estado de Tocantins, com os dados do INEP. Na exploração dos resultados da pesquisa, foram utilizadas soluções de software e técnicas de estatística em mineração de dados (SAS e ANOVA), juntamente com planilhas eletrônicas. A análise identificou atributos dos indicadores com significância estatística para as escolas que tiveram melhor desempenho na Prova Brasil.

Os autores Silva et.al (SANTOS; MEDEIROS, 2020) utilizaram a base do IDEB e dados do Portal da Transparência de financiamento federal da educação básica do estado da Paraíba visando estudar se havia uma relação do investimento federal com a qualidade da educação nos anos 2016 e 2017. Eles utilizaram a metodologia KDD, e os experimentos foram realizados no software R, recorrendo a dois algoritmos: Regressão Linear Simples e Correlação de Pearson. As notas do IDEB foram divididas em duas categorias: anos iniciais e anos finais. A regressão realizada não mostrou resultado significativo, portanto, não foi encontrada linearidade na relação entre o investimento Federal e o IDEB. Os resultados da correlação de Pearson se mostraram também pouco significativas, uma vez que o coeficiente de correlação resultante, respectivamente, para o IDEB anos iniciais e finais foram de 0.01 e 0.05. Segundo o que os pesquisadores concluíram, não há relação direta entre o investimento federal e o IDEB. Eles concluíram que somente os dados de financiamento federal na educação municipal são pouco representativos ou não são suficientes para obter uma relação de causalidade entre o investimento federal bianual e a média do IDEB para os municípios.

Os pesquisadores Silva et al. (SILVA; SOUZA; CYSNEIROS, 2021) utilizaram dados nacionais do SAEB para a elaboração de uma ferramenta para extração de conhecimento através de análise de dados poligonais simbólicos (PSDA). Trata-se de uma estrutura que extrai conhecimento usando polígono regular formado a partir de dados em classe, big data e dados complexos. Para a elaboração da ferramenta, foram usadas as principais

medidas descritivas das variáveis, por exemplo, média, variância, correlação, e também é realizado um modelo de regressão linear poligonal. O experimento foi realizado através do software R. Primeiramente, foi realizada a extração de informações a partir da base de dados, formando-se uma nova base, sem dados discrepantes ou ausentes. Em seguida, foi aplicado o pacote PSDA através do software R, com a disponibilização de medidas descritivas, gráficos de dispersão e um modelo de regressão para dados poligonais.

2.3 Análise Comparativa e Discussão

A Tabela 1 apresenta uma síntese dos trabalhos relacionados. A primeira coluna da tabela apresenta as referências dos artigos. A segunda coluna apresenta as bases de dados utilizadas (e.g., SAEB, IDEB, Censo Escolar e Indicadores Educacionais), enquanto a terceira coluna apresenta o problema estudado em cada trabalho (e.g, Baixo Desempenho Escolar, Evasão de Alunos, Falta de Padrão entre Bases de Dados Educacionais, Gestão de Baixa Qualidade, Baixo Investimento na Educação). A última coluna mostra a região que se refere os dados minerados.

A maioria dos trabalhos encontrados através da RSL utiliza dados gerais de todo Brasil, chamados de dados nacionais. Alguns trabalhos relacionados destacam-se por utilizarem dados regionalizados e específicos de estados ou cidades. Dessa forma, esta pesquisa de dissertação se assemelha a estes por usar dados especificamente regionalizados, sendo, para o melhor do nosso conhecimento, o primeiro estudo a utilizar os dados do estado do Maranhão.

De acordo com os artigos obtidos na RSL, o principal problema estudado é o baixo desempenho escolar, o qual refere-se, por exemplo, ao baixo desempenho dos estudantes no SAEB e IDEB ou ENEM. O INEP fornece todos os anos uma estimativa sobre as notas do IDEB de acordo com os dados do ano anterior, desejando que as escolas alcancem aquelas notas para melhorar o nível da educação básica brasileira. É tácito que algumas escolas não alcançam essa estimativa. Dessa forma, o baixo desempenho escolar foi um problema frequente tratado nos artigos revisados. Também foram encontrados outros problemas como a fragilidade do sistema educacional, evasão e reprovação de alunos, problemas com a formação dos docentes, baixo investimento na educação e dois artigos trataram da falta de padronização das bases de dados educacionais.

As contribuições desta pesquisa estão voltadas para a aplicação de técnicas de mineração de dados, através da análise de correlação e análise de fatores, e elaboração de modelos exploratórios de árvore de decisão e regressão linear, com os dados do SAEB e IDEB do ensino médio público do Maranhão. O trabalho pode contribuir com evidências e geração de *insights* para a tomada de decisões aplicadas na realização de políticas públicas educacionais de forma que possa melhorar as notas do IDEB. O trabalho também

Tabela 1 – Caracterização dos trabalhos relacionados.

Referência	**Dados**	**Problema**	**Região**
(SOARES, 2006)	SAEB	Baixo desempenho escolar	Nacional
(WALTENBERG; VANDENBERGHE, 2007)	SAEB	Baixo desempenho escolar	Nacional
(RAMOS; MACHADO; CORDEIRO, 2015)	IDEB e SAEB	Fragilidade do sistema educacional	Nacional
(PENTEADO, 2016a)	SAEB	Melhoria de desempenho de alunos	Nacional
(PENTEADO, 2016b)	IDEB	Baixo desempenho escolar	Nacional
(JÚNIOR *et al.*, 2017)	Censo	Baixo desempenho escolar	Pernambuco
(CARVALHO; CRUZ; GOUVEIA, 2017)	Censo	Evasão de alunos	Pernambuco
(BEM; PEREIRA; SOUZA, 2017)	IDEB	Falta de padronização entre bases educacionais	Pajeú - PE
(AVILA *et al.*, 2018)	Indicadores	Falta de padronização entre bases educacionais	Nacional
(NASCIMENTO; JÚNIOR, 2018)	Indicadores	Formação docente de baixa qualidade	Nacional
(NASCIMENTO; JUNIOR; FAGUNDES, 2018)	Indicadores	Evasão e reprovação de alunos	Nacional
(NASCIMENTO; FAGUNDES; MACIEL, 2019)	Indicadores	Aproximação do desempenho previsto das taxas de eficiência	Nacional
(PINTO; JÚNIOR; COSTA, 2019)	SAEB	Baixo Desempenho escolar	T. Vilela - AL
(JUNIOR *et al.*, 2019)	Indicadores	Reprovação de alunos	Nacional
(JÚNIOR *et al.*, 2019)	IDEB e SAEB	Gestão de baixa qualidade e baixo desempenho escolar	Maceió - AL
(PINTO *et al.*, 2019)	SAEB	Baixo desempenho escolar	Maceió - AL
(PACINI, 2020)	SAEB	Baixo desempenho escolar	Tocantins
(SANTOS; MEDEIROS, 2020)	IDEB	Baixo investimento na educação	Paraíba
(SILVA; SOUZA; CYSNEIROS, 2021)	SAEB e Censo	Qualidade do processo de mineração de dados	Nacional
Este Trabalho	SAEB e IDEB	Baixo desempenho escolar	Maranhão

destaca a relevância da mineração de dados como ferramenta importante na descoberta de conhecimento para a realização de políticas públicas educacionais baseadas em evidência.

3 Entendimento do Negócio e dos Dados

Esse capítulo apresenta, primeiramente, o entendimento da educação do Maranhão, e posteriormente o entendimento dos dados. Na seção seguinte, ele descreve as ferramentas computacionais utilizadas. Por fim, a última seção deste capítulo apresenta uma conclusão das atividades realizadas nestas duas etapas do estudo.

As duas primeiras etapas da metodologia CRISP-DM usada neste trabalho são o entendimento do negócio e o entendimento dos dados. A etapa de entendimento do negócio consiste na proposição de uma situação-problema relacionada à mineração de dados e um plano preliminar para solucionar esse problema. Nesta fase, são definidos os objetivos do projeto e os recursos a serem utilizados na mineração. Já a etapa de entendimento dos dados consiste em buscar a base de dados e fazer a análise preliminar deles para verificar se é possível resolver o problema abordado, ou pelo menos parte dele. Nessa etapa também já é feita uma análise preliminar dos dados.

3.1 Entendendo a Educação Básica Maranhense

O estado do Maranhão está localizado na região Nordeste do Brasil. De acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE), a população estimada é de 7 milhões de habitantes distribuídas em 216 cidades, com uma densidade demográfica de 19 habitantes por quilômetro quadrado. A renda domiciliar per capita é de R$ 636,00 (IBGE, 2021). Ainda segundo dados do IBGE de 2010, o Índice de Desenvolvimento Humano (IDH) é 0,639 (IBGE, 2021). Esse número coloca o estado do Maranhão na 26ª posição entre os estados brasileiros e o Distrito Federal (DF). Esse dado também demonstra o estado do Maranhão como um dos mais pobres e menos desenvolvidos dentre os estados brasileiros.

Segundo dados do Censo Escolar de 2019, o estado do Maranhão possui 1.031 escolas de ensino médio, sendo: 28 escolas federais, 14 escolas municipais, 193 escolas privadas e 796 escolas estaduais. O Maranhão constitui, portanto, a rede estadual com o maior número de escolas do Brasil.

No que diz respeito ao IDEB, os números do Maranhão, referente às escolas públicas de ensino médio, não tiveram uma grande variação nas últimas avaliações ((SEDUC-MA), 2020). As médias das últimas avaliações foram: 2,8 em 2013, 3,1 em 2015, 3,4 em 2017, e 3,7 em 2019, como visto na Tabela 2. Considerando o IDEB em uma escala de 0 (zero) a 10 (dez), as médias referentes às escolas públicas de ensino médio do Maranhão estão muito aquém de uma nota considerada elevada para os padrões nacionais. Isso evidencia a

necessidade de verificação dos fatores que influenciam essas notas do IDEB, que é o objeto de estudo desta dissertação de mestrado.

Tabela 2 – Resultados do IDEB do ensino médio da rede pública estadual dos estados brasileiros referente ao ano 2019 (INEP, 2020b).

Estado	***IDEB 2019***	***Posição***
Goiás	4,7	1º
Espirito Santo	4,6	2º
Paraná	4,4	3º
Pernambuco	4,4	3º
São Paulo	4,3	4º
Ceará	4,2	5º
Mato Grosso do Sul	4,1	6º
Distrito Federal	4,0	7º
Minas Gerais	4,0	7º
Rio Grande do Sul	4,0	7º
Santa Catarina	4,0	7º
Tocantins	3,9	8º
Santa Catarina	3,8	9º
Acre	3,7	10º
Maranhão	**3,7**	10º
Piauí	3,7	10º
Alagoas	3,6	11º
Paraíba	3,6	11º
Amazonas	3,5	12º
Rio de Janeiro	3,5	12º
Roraima	3,5	12º
Mato Grosso	3,4	13º
Sergipe	3,3	14º
Amapá	3,2	15º
Bahia	3,2	15º
Pará	3,2	15º
Rio Grande do Norte	3,2	16º

No estado do Maranhão, além das tradicionais escolas públicas de ensino médio, o Estado recentemente criou escolas de ensino técnico e em tempo integral. As escolas técnicas criadas pelo Governo do Maranhão são chamadas de Institutos Estaduais de Educação, Ciência e Tecnologia do Maranhão (IEMAs) e têm o objetivo de ofertar cursos técnicos integrados ao ensino médio. Já as escolas de tempo integral são chamadas Centros Educa Mais. Elas oferecem educação em tempo integral, em que o aluno passa o dia (de 7h às 17h) na escola, fazendo refeição, e tendo horas de estudo e lazer na própria escola. Os IEMAs e os Centros Educa Mais já estão presentes em 33 cidades.

Atualmente, o piso salarial nacional dos professores da educação básica é de R$ 3.845,63 (2022), referente a jornada semanal de 40 horas. No Maranhão, o salário do professor de ensino médio em início de carreira é de R$ 6.867,68 (MARANHãO, 2022), pela jornada semanal de 40 horas. Esse valor do salário de professores no Maranhão é composto de salário-base e mais Gratificação de Atividade do Magistério (GAM), pago a professores efetivos concursados. Há, também, os professores concursados com jornada semanal de 20 horas, cujo valor do salário é proporcional ao valor de 40 horas. O Maranhão também conta com professores contratados temporariamente para suprir as necessidades de carência de professores. Atualmente, o professor contratado no Maranhão recebe salário bruto mensal de R$ 1.876,06 pela jornada semanal de 20 horas.

A nota do IDEB é usada para medir a qualidade da educação. Dessa forma, quanto menor a nota, menor a qualidade educacional, caracterizando uma educação com desempenho ruim. A baixa nota do IDEB do estado do Maranhão pode ser causada por vários motivos que precisam ser investigados e estudados. As abordagens de investigação podem ser de diversas maneiras. Nesta pesquisa, o problema de estudo é não haver uma abordagem do ponto de vista da mineração de dados, e considerando os fatores que podem causar impacto e influência na qualidade educacional da rede pública do estado Maranhão.

A seguir, como parte da etapa de entendimento do negócio, definimos o objetivo do processo de mineração de dados, as questões de pesquisa e os critérios de sucesso.

Definição do objetivo: O objetivo deste estudo, ou seja, do processo de mineração de dados a ser realizado, é entender os fatores que influenciam na qualidade educacional do estado do Maranhão. Para isso, o estudo tira vantagem do uso de técnicas de mineração das bases de dados fornecidas pelo INEP: a base do SAEB e a base do IDEB.

A pesquisa busca responder às seguintes **Questões de Pesquisa (QPs)**:

- (QP1) Quais fatores mais influenciam o desempenho acadêmico das escolas de ensino médio do estado do Maranhão?
- (QP2) É possível predizer a nota do IDEB a partir de um conjunto de fatores?

Como **Critérios de Sucesso (CSs)** dessa pesquisa, podemos listar:

- (CS1) O cumprimento dos objetivos almejados e a identificação das respostas para as questões de pesquisa;
- (CS2) A identificação do estado da arte do assunto estudado através da elaboração de uma RSL cujos detalhes constam no Capítulo 2;
- (CS3) A análise preliminar dos dados apresentados nas bases do SAEB e IDEB;
- (CS4) A identificação e análise dos fatores obtidos a partir da mineração dos dados;
- (CS5) A realização de publicação de artigos científicos para realização da etapa de distribuição da metodologia CRISP-DM.

3.2 Entendimento dos Dados do IDEB e SAEB do Maranhão

Existem diversas bases de dados educacionais brasileiras disponibilizadas pelo INEP e, dentre elas, estão o Censo Escolar, o SAEB, o ENEM, o Exame Nacional para Certificação da Educação de Jovens e Adultos (ENCCEJA), o Exame Nacional de Desempenho de Estudantes da Educação Superior (ENADE), o Censo da Educação Superior e Indicadores Educacionais. As bases SAEB e IDEB são utilizadas nesta pesquisa e, por esse motivo, são detalhadas a seguir.

As avaliações aplicadas pelo SAEB acontecem a cada dois anos. As provas de duas disciplinas são aplicadas, Língua Portuguesa e Matemática, para estudantes do 5º ano do Ensino Fundamental I, 9º ano do Ensino Fundamental II, e 3º ano do Ensino Médio.

O IDEB é um indicador que mede a qualidade da educação básica brasileira, calculado a partir da média da proficiência das provas de Língua Portuguesa e Matemática aplicadas pelo SAEB, padronizada entre 0 (zero) e 10 (dez), multiplicada pelo indicador de rendimento baseado na taxa percentual de aprovação dos alunos, a qual é padronizada entre 0 (zero) e 1 (um). Além das avaliações, os questionários do SAEB também são aplicados aos alunos, professores e gestores da educação. Eles visam obter dados sobre a infraestrutura das escolas, aspectos administrativos e socioeconômicos.

Todos os dados produzidos por meio de avaliações e questionários são disponibilizados pelo INEP através de uma página para acesso a dados abertos (INEP, 2020a). O volume de dados disponível é grande e tem sido analisado para gerar conhecimento para gestores e demais envolvidos na educação. Isso permite a realização de tomadas de decisão baseadas em evidências. Considerando essa disponibilização de bases, pesquisas em mineração de dados educacionais têm sido conduzidas, como, por exemplo, as levantadas na RSL (Apêndice A).

Para esta pesquisa, foram analisados os dados de escolas públicas estaduais de ensino médio do Maranhão, obtidos através das bases de dados do SAEB e IDEB referentes ao ano de 2019 por ser os dados mais recentes disponíveis.

Quando se realiza uma análise de dados, as informações podem vir de uma variedade de fontes e em vários formatos. Nesta etapa, os dados foram compreendidos, e a qualidade e adequação desses dados foram verificadas. Nessa etapa, também foram feitas as primeiras constatações e questionamentos baseados nos dados.

O primeiro conjunto de dados utilizado foi o que continha as respostas do questionário socioeconômico aplicado aos alunos do SAEB. Ele é um conjunto com dados apenas de alunos do 3º ou 4º ano do ensino médio de todo o Brasil. São mais de 2 milhões de linhas, que correspondem aos dados de cada aluno que respondeu ao questionário; e mais de 90 variáveis, entre elas:

- Dados de identificação: código da escola, região, estado, cidade;
- Dados das respostas da Prova Brasil: respostas do questionário socioeconômico com informações como raça, meio de transporte utilizado para ir à escola, escolaridade dos pais, questões sobre incentivo dos pais para estudar, acesso à tecnologia.

O segundo conjunto de dados analisado contém o resultado do IDEB para o ano de 2019. Ele contém mais de 20 mil registros, que correspondem a cada uma das escolas participantes do SAEB, e 28 variáveis, entre elas estão os dados de identificação da escola, como código INEP, estado, região, cidade. Além de analisar o conjunto de dados, o INEP também disponibiliza um dicionário em formato XLS, com descrição detalhada de todos os componentes do conjunto de dados.

Todos os dados utilizados da pesquisa tiverem como fonte a página de dados abertos do INEP (INEP, 2020a). Trata-se de um portal onde são disponibilizadas bases de dados referentes a cada ano em que os dados são coletados.

As duas bases de dados utilizadas possuem qualidade e podem ser consideradas idôneas. A justificativa para isso é que são obtidos a partir de fontes oficiais. As duas bases também contêm dados faltantes. Na base de dados do SAEB, os dados faltantes são dos alunos que não responderam ao questionário socioeconômico. Essa identificação é possível, pois há uma variável com resposta binária em que informa se o aluno respondeu ou não ao questionário.

Quanto aos dados do IDEB, os dados ausentes se referem às escolas que por algum motivo não realizaram a prova para obtenção da nota ou não tiveram quantidade de alunos suficiente para que a nota fosse divulgada. Os dois conjuntos de dados também possuem dados nulos. Esses conjuntos de dados nulos são recorrentes de preenchimento com caracteres fora da resposta padrão esperada.

A Tabela 3 apresenta a lista de variáveis (i.e., as questões) e a respectiva descrição, e também as respostas para cada questão.

Tabela 3 – Descrição das variáveis utilizadas.

Variável	**Descrição**	**Resposta**
TEM COMPUTADOR	O aluno tem computador em casa?	A. Nenhum; B. um; C. dois; D. três ou mais
TEM WIFI	O aluno tem acesso à rede de internet sem fio em casa?	a. Sim; B. Não
ESCOL MAE	Nível de escolaridade da mãe	A. Não completou o 5º ano do Ensino Fundamental. B. Ensino Fundamental, até o 5º ano. C. Ensino Fundamental completo. D. Ensino Médio completo. E. Ensino Superior completo (faculdade ou graduação). F. Não sei.
ESCOL PAI	Nível de escolaridade do pai	A. Não completou o 5º ano do Ensino Fundamental. B. Ensino Fundamental, até o 5º ano. C. Ensino Fundamental completo. D. Ensino Médio completo. E. Ensino Superior completo (faculdade ou graduação). F. Não sei.
PAIS CONVERSAM ESCOLA	Os pais do aluno conversam com ele sobre a escola?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre
PAIS INCENTIVAM	Os pais incentivam o aluno a estudar?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre
PAIS TAREFA CASA	Os pais incentivam o aluno a fazer as tarefas de casa?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre
INCENTIVAR IR ESCOLA	Os pais incentivam o aluno a ir para a escola?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre
COMPARECER REUNIOES	Os pais comparecem às reuniões com os professores e gestão da escola?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre
LER NOTICIAS	Os alunos lêm noticias?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre
LER LIVROS	Os alunos lêm livros	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre
LER QUADRINHOS	Os alunos lêm revistas em quadrinhos?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre
DEVER CASA	Os alunos fazem o dever de casa?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre
IDADE INIC ESTUDAR	Com que idade os alunos começaram a estudar?	A. 3 anos ou menos. B. 4 ou 5 anos. C. 6 ou 7 anos. D. 8 anos ou mais.
REPROVOU	O aluno já reprovou algum ano durante sua trajetória escolar?	B. Nunca; B. Sim, uma vez; C. Sim, duas vezes ou mais;
EVADIU	O aluno já evadiu da escola durante algum ano da sua trajetória escolar?	B. Nunca; B. Sim, uma vez; C. Sim, duas vezes ou mais;
IDEB 2019	Nota do IDEB referente ao ano de 2019	Notas de 0 a 10

3.3 Ferramentas Utilizadas

O entendimento dos dados é uma etapa que também requer o planejamento de como atingir o objetivo definido. Isso inclui o levantamento de estruturas, como as tecnologias

computacionais que serão utilizadas. A realização desta pesquisa exigiu a utilização de diversas soluções tecnológicas. Desde linguagem de programação e suas bibliotecas, passando por aplicações web e soluções de software desktop. Esta seção descreve as tecnologias utilizadas nesta pesquisa.

A linguagem de programação **Python** (PYTHON, 2020) é bastante usada em Ciência de Dados. Atualmente está na versão 3.9.1, com uma sintaxe que permite a escrita de comandos em poucas linhas e de forma simplificada. A utilização de *Python* vai muito além, desde seu uso no desenvolvimento web, em *frameworks* de programação e sistemas de gerenciamento de conteúdos. Bibliotecas de *Python* podem suportar diversos protocolos como HTML, XML, JSON. *Python* é usada na computação científica e numérica e no âmbito educacional é usada para iniciação à programação por ser fácil de usar e de aprender. A linguagem pode ser executada em compiladores web pelo navegador ou instalados localmente.

Pandas é uma biblioteca escrita na linguagem Python (PANDAS, 2020) especificamente para análise de dados, pois fornece ferramentas para manipulação de estruturas de dados de forma simples. Ela permite a execução de operações complexas que utilizam matrizes e vetores com ótimo desempenho. A biblioteca permite a manipulação de dados e estruturas de dados na memória em diferentes formatos: arquivos CSV e de texto, Microsoft Excel, bancos de dados SQL e o formato rápido HDF5. Dentre algumas possibilidades de manipulação de conjuntos de dados únicos, podemos citar:

- As colunas podem ser inseridas e excluídas de estruturas de dados para mutabilidade de tamanho;
- Operações dividir-aplicar-combinar em conjuntos de dados;
- Mesclagem e junção de conjuntos de dados de alto desempenho.

A biblioteca **NumPy** (NUMPY, 2020), escrita na linguagem Python, permite implementação de *arrays* multidimensionais e com a fácil execução de operações matemáticas e lógicas. Por exemplo: ordenação, seleção, transformações, operações estatísticas. NumPy é um projeto de código aberto criado em 2005, com base no trabalho inicial das bibliotecas *Numérical* e *Numarray*. A biblioteca é de código aberto e desenvolvida abertamente na plataforma *GitHub* pela comunidade científica. Devido ao grande crescimento da biblioteca, atualmente o projeto dispõe de equipes específicas para desenvolvimento de código, documentação, *website*, triagem, financiamento e subsídios e administração.

Matplotlib é uma biblioteca Python utilizada para a visualização e plotagem de gráficos (MATPLOTLIB, 2020). Pode ser empregada na geração de diversos tipos de gráficos como histogramas, gráficos de barras, gráficos de pizza, de forma fácil e rápida.

Criada pelo biólogo e neurocientista americano John D. Hunter, possui uma comunidade ativa de desenvolvedores, é distribuída sob uma licença BSD. A biblioteca também permite a criação e exibição de mapas, pontos arbitrários, linhas e polígonos, com capacidades para transformação de imagem e também gráficos 3D. Outras funcionalidades podem ser adicionadas através de aplicações de terceiros baseadas na biblioteca Matplotlib, como *Seaborn, HoloViews, Ggplot*, e um kit de projeção e mapeamento denominado *Cartopy*.

A biblioteca **Scikit-Learn** é escrita na linguagem Python e utilizada para trabalhar com *Machine Learning* (Aprendizado de Máquina) (SCIKIT-LEARN, 2020). Ela contém diversos algoritmos implementados, que permitem a realização de métodos de análise e processamento de dados, criação de modelos e realização de avaliações com diversas métricas.

O **Google Colaboratory** (Colab) é uma aplicação web de propriedade do Google que permite escrever códigos Python sem fazer configuração nem instalação local, com acesso grátis e fácil compartilhamento (COLABORATORY, 2020). Semelhante à aplicação *Jupyter*, a interface também é apresentada em células, porém o *Colab* não tem versão para instalação na máquina local. Pode ser usada diversas linguagens de programação, e ao recorrer à linguagem Python para ciência de dados, é possível importar as bibliotecas específicas para esta finalidade. A ferramenta permite importar um conjunto de dados de imagem, treinar um classificador e avaliar o modelo. Os códigos são executados nos servidores em nuvem do Google, o que permite o maior poder de processamento com recursos que podem ir além dos disponíveis em uma máquina local, como a utilização de *Graphics Processing Units* (GPUs) e *Tensor Processing Units* (TPUs). A ferramenta também é integrada ao Google Drive, permitindo o armazenamento e uso de conjuntos de dados armazenados na nuvem.

O **Orange** é um software de código aberto em que é possível a visualização de dados por meio de sua interface simplificada e intuitiva, com apenas alguns cliques (ORANGE, 2020). É possível extrair dados via programação visual ou *scripts* Python, explorar estatísticas, realizar *box plots* ou *scatter plots* e aprofundar dados com árvores de decisão, agrupamento hierárquico, *heatmaps* e projeções lineares. A ferramenta possui também complementos de mineração de dados de fontes externas para execução de processamento de linguagem natural, mineração de texto, bioinformática, e mineração de regras de associação. O Orange foi utilizado nesta pesquisa para elaboração da árvore de decisão por sua facilidade de manuseio e uso.

O **IBM SPSS** (IBM, 2022) é um software utilizado para fazer análise estatística. Ele apresenta uma interface amigável, fácil de usar, e um conjunto e funcionalidades para fazer a análise estatísticas da base de dados. Além disso, ao inserir uma base de dados, antes de fazer as análises, também é possível, caso seja necessário, realizar um pré-processamento para esses dados para quem fiquem prontos para serem trabalhados. O

software foi utilizado na etapa de identificação e análise de fatores. A base de dados foi inserida já totalmente pronta, portanto a preparação foi realizada no *Colab*, não sendo necessário realizar preparação de dados no SPSS, justificando-se assim sua escolha. O uso do **IBM SPSS** justifica-se nesta pesquisa por ser um software com funcionalidad3s específicas e relacionadas ao objetivo do trabalho proposto.

3.4 Conclusão

As duas primeiras etapas da metodologia foram realizadas com sucesso. O entendimento do negócio (i.e., educação maranhense) foi realizado a partir do estado de conceitos relacionados ao tema, e o entendimento dos dados também foi realizado. A partir de então, elaborado o objetivo da pesquisa e o plano inicial para conseguir alcançá-lo. As próximas etapas envolvem a preparação dos dados para ficarem prontos para a realização da etapa de modelagem.

4 Preparação e Modelagem

Este capítulo apresenta a preparação dos dados do SAEB e do IDEB e os resultados para as questões de pesquisa. Primeiramente, é apresentado como os dados foram preparados, em seguida os modelos e resultados obtidos após os dados estarem prontos para uso: análise de correlação, análise de fatores, modelos de regressão e árvore de decisão.

4.1 Pré-processamento

Esta pesquisa realiza uma mineração de dados educacionais fornecidos pelo INEP. As bases fornecem dados educacionais de todos os estados do Brasil, sendo possível filtrar e separá-los por regiões (estados ou municípios), por dependência administrativa (municipal, estadual, federal e privada), por nível de ensino (fundamental e médio), conforme os dados disponíveis. Para esta pesquisa, foram utilizados dados de escolas públicas de ensino médio da rede estadual do Maranhão, pois a proposta é estudar o que impacta os indicadores de qualidade da educação pública maranhense.

A etapa de preparação pode envolver, caso necessário, a padronização e normalização dos dados, para produzir o conjunto final de dados, pronto para ser analisado. Os principais dados da base do SAEB são as respostas do questionário socioeconômico e da prova (respostas, notas, médias, dentre outros). Como nessa base de dados o que interessa são as respostas do questionário socioeconômico, a primeira filtragem feita foi para deixar apenas as linhas com os alunos que haviam respondido o questionário socioeconômico, resultando num total de 49 mil registros.

Com o conjunto de dados apresentando somente dados de escolas do Maranhão, excluem-se os dados de escolas federais e particulares, restando apenas as escolas públicas estaduais. Em seguida, o conjunto de dados foi alterado para exibir apenas as variáveis a serem trabalhadas. Embora o arquivo possua mais de 90 variáveis, para fins da pesquisa, neste primeiro conjunto de dados foram utilizadas 16 variáveis, sendo de interesse do estudo.

A base de dados contem dados categóricos e dados numéricos. Por esse motivo os dados tiveram que ser padronizados e normalizados. Os dados categóricos foram convertidos para dados numéricos.

O conjunto de dados com os resultados do IDEB também precisava ser tratado. Assim como os dados do aluno, ele contém diversas variáveis. Como o que nos interessa, nessa pesquisa, é apenas a variável dos resultados, deixamos apenas esta e a variável com o código das escolas, pois é necessário juntar os conjuntos de dados. Todas as linhas que

não possuíam dados do IDEB também foram excluídas, pois, o IDEB de algumas escolas não constam no conjunto de dados com os resultados.

Em seguida, foi realizado a junção das duas bases de dados. Ou seja, elas foram mescladas em uma única base adicionando ao conjunto de dados do SAEB uma coluna com as notas do IDEB para o ano de 2019. Todo o processo realizado na etapa de preparação dos dados é apresentado na Figura 3. As Tabelas 4 e 5 sumarizam (funções *describe* e *info*) o conjunto de dados utilizados no estudo após a realização da etapa de preparação dos dados.

Tabela 4 – Descrição do conjunto de dados final utilizado no estudo.

Variável	***Count***	***mean***	***std***	***min***	***25%***	***50%***	***75%***	***max***
TEM COMPUTADOR	27595	0.45	0.69	0.0	0.0	0.0	1.0	3.0
TEM WIFI	27595	0.51	0.49	0.0	0.0	1.0	1.0	1.0
ESCOL MAE	27595	3.13	1.33	1.0	2.0	4.0	4.0	5.0
ESCOL PAI	27595	2.74	1.34	1.0	1.0	3.0	4.0	5.0
PAIS CONVERSAM ESCOLA	27595	2.24	0.62	1.0	2.0	2.0	3.0	3.0
PAIS INCENTIVAM	27595	2.82	0.436	1.0	3.0	3.0	3.0	3.0
PAIS TAREFA CASA	27595	2.50	0.66	1.0	2.0	3.0	3.0	3.0
INCENTIVAR IR ESCOLA	27595	2.87	0.41	1.0	3.0	3.0	3.0	3.0
COMPARECER REUNIOES	27595	2.47	0.66	1.0	2.0	3.0	3.0	3.0
LERNOTICIAS	27595	2.20	0.61	1.0	2.0	2.0	3.0	3.0
LERLIVROS	27595	2.07	0.64	1.0	2.0	2.0	2.0	3.0
LERQUADRINHOS	27595	1.76	0.70	1.0	1.0	2.0	2.0	3.0
DEVER CASA	27595	2.04	0.91	0.0	1.0	2.0	3.0	3.0
IDADE INIC ESTUDAR	27595	2.40	0.71	0.0	2.0	3.0	3.0	3.0
REPROVOU	27595	2.61	0.61	1.0	2.0	3.0	3.0	3.0
EVADIU	27595	2.90	0.35	1.0	3.0	3.0	3.0	3.0
IDEB 2019	27595	3.82	0.67	1.7	3.4	3.7	4.2	6.20

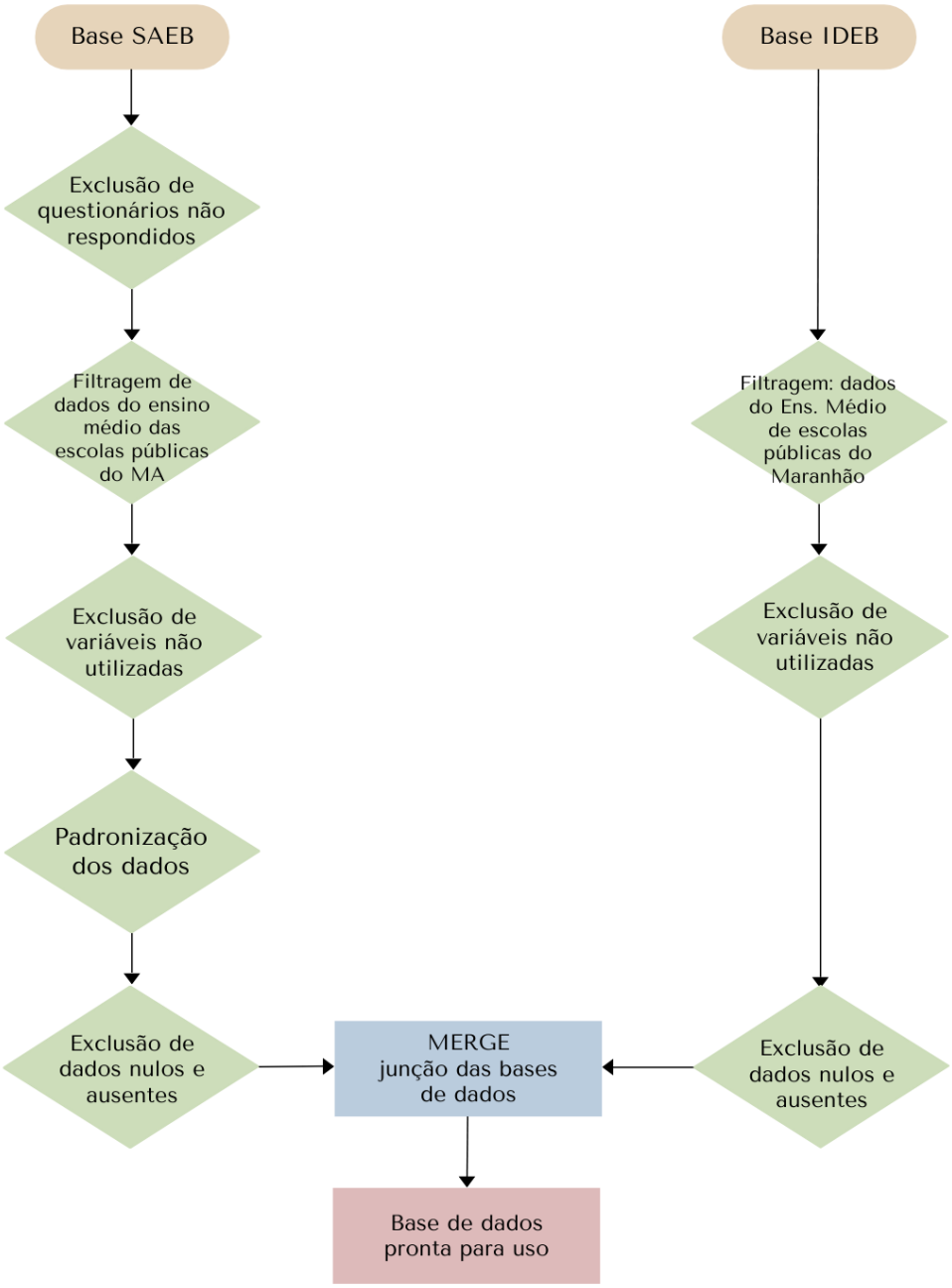

Figura 3 – Organização da etapa de preparação dos dados.

4.2 Análise Descritiva

Com os dados prontos, foi possível realizar uma análise descritiva para efeito de estudo e observação do comportamento dos dados e das variáveis disponíveis. A Figura 4 mostra o grau de escolaridade das mães e dos pais. Ambos possuem como escolaridade predominante o ensino médio completo. No entanto, a quantidade de mães é bem maior que a de pais. Isso levanta uma hipótese de que há muitos alunos que convivem exclusivamente com a mãe, sem a presença do pai. No entanto, essa hipótese até o momento não pôde ser

Tabela 5 – Informações do conjunto de dados final utilizado no estudo.

	Column	***Non-Null Count***	***Dtype***
0	TEM COMPUTADOR	27595 non-null	int64
1	TEM WIFI	27595 non-null	int64
2	ESCOL MAE	27595 non-null	int64
3	ESCOL PAI	27595 non-null	int64
4	PAIS CONVERSAM ESCOLA	27595 non-null	int64
5	PAIS INCENTIVAM	27595 non-null	int64
6	PAIS TAREFA CASA	27595 non-null	int64
7	INCENTIVAR IR ESCOLA	27595 non-null	int64
8	COMPARECER REUNIOES	27595 non-null	int64
9	LERNOTICIAS	27595 non-null	int64
10	LERLIVROS	27595 non-null	int64
11	LERQUADRINHOS	27595 non-null	int64
12	DEVER CASA	27595 non-null	int64
13	IDADE INIC ESTUDAR	27595 non-null	int64
14	REPROVOU	27595 non-null	int64
15	EVADIU	27595 non-null	int64
16	IDEB 2019	27595 non-null	float64

comprovada através desta análise.

A Figura 5 ilustra a quantidade de computadores que os alunos possuem e o acesso à Internet através de rede sem fio. A maioria dos alunos não possui nenhum computador. No entanto, mais da metade dos alunos possuem acesso à rede Wi-Fi.

O IDH-M mede o desenvolvimento das cidades considerando três dimensões: longevidade, educação e renda. A Figura 6 ilustra, através de um *scatter plot*, a distribuição do IDEB em função do IDH-M das cidades maranhenses. A Figura mostra que quanto menor o IDH-M, há uma tendência do IDEB ser abaixo da média, com número maior de cidades nesse primeiro quadrante (cor púrpura). O segundo quadrante (verde) ilustra uma menor quantidade de cidades com IDEB acima da média em função do IDH-M baixo. O terceiro quadrante (azul), embora tenha um número considerável de cidades com IDEB acima da média em função do alto IDH-M, não chega a ser maior que o primeiro quadrante. Por fim, o quadrante de cor laranja mostra poucas cidades com bom desempenho no IDH-M, porém com IDH-M abaixo da média.

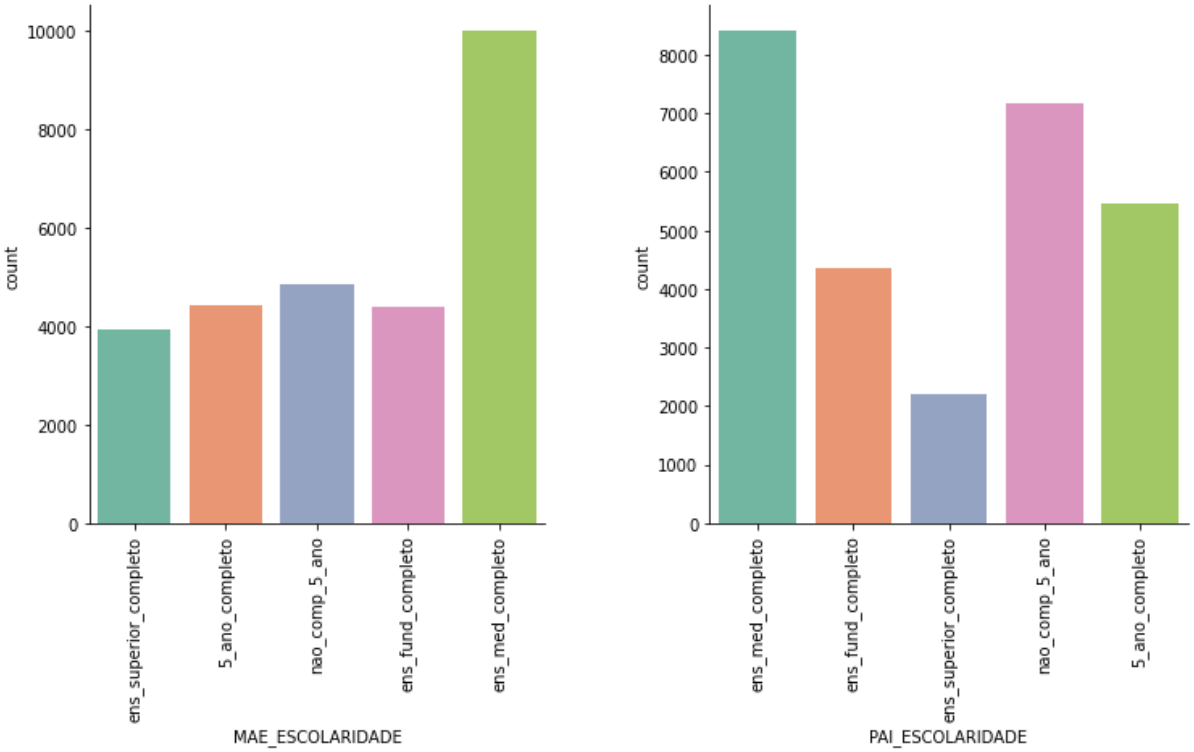

Figura 4 – Escolaridade das mães e dos pais.

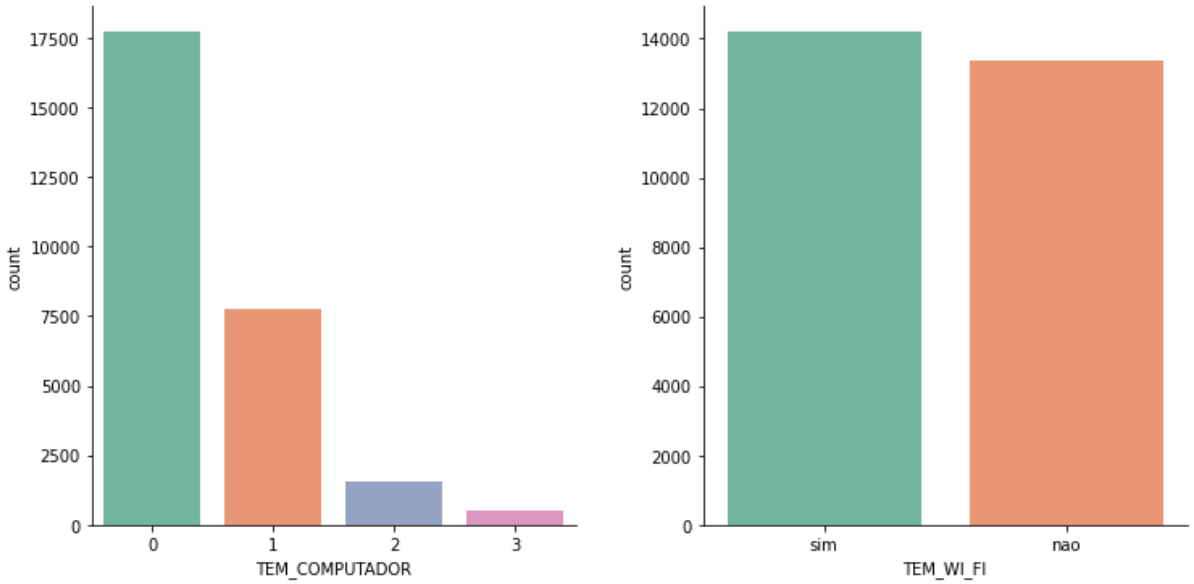

Figura 5 – Recursos tecnológicos: quantidade de computadores que cada aluno possui e acesso à Internet sem fio.

4.3 Análise de Correlação

4.3.1 Conceitos

Correlação é uma medida estatística que indica a variação conjunta de duas ou mais variáveis. A correlação pode ser positiva ou negativa. Se a correlação for positiva, ela se refere ao aumento ou diminuição das variáveis conjuntamente. No entanto, se a

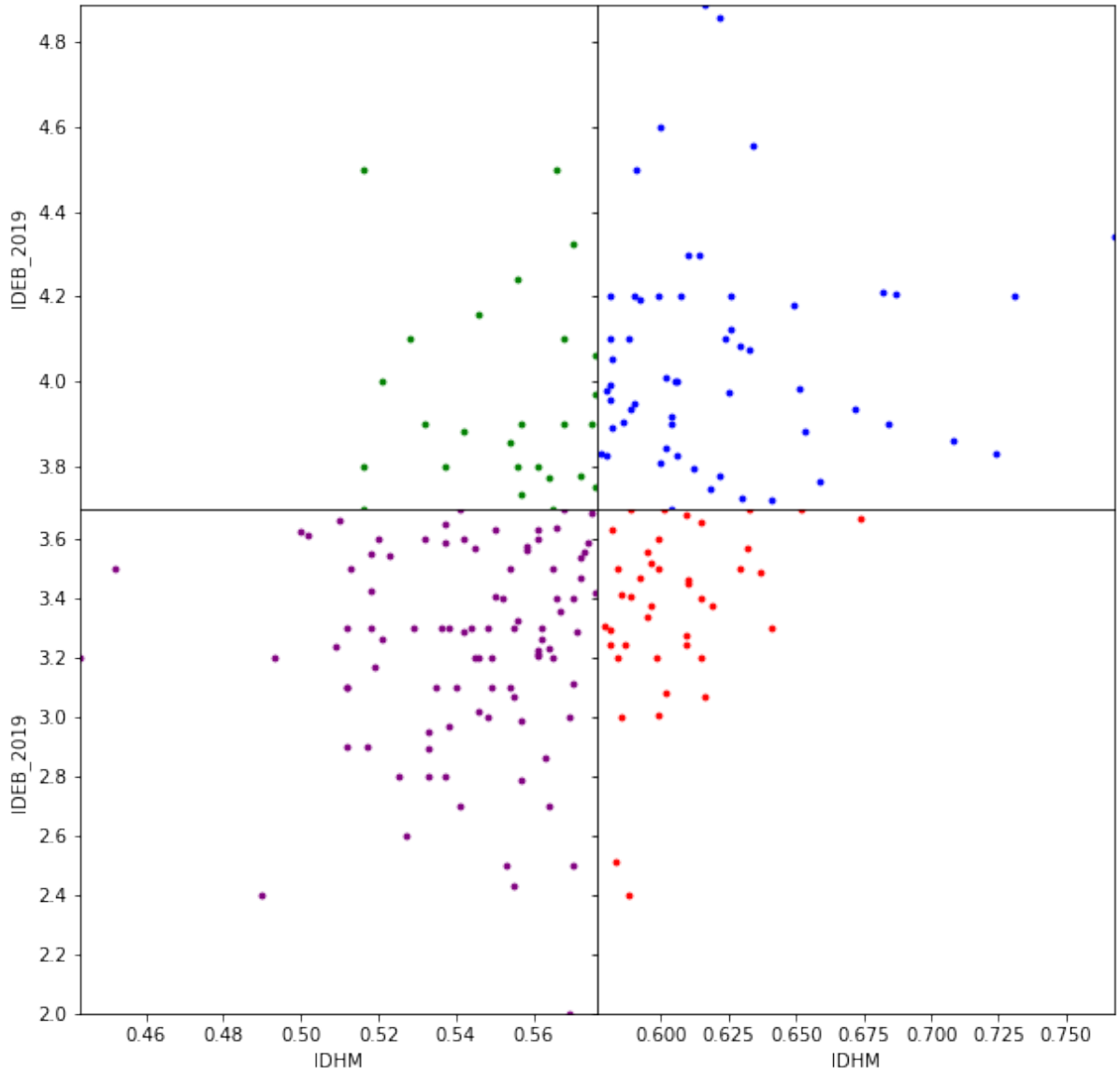

Figura 6 – *Scatter plot* das variáveis IDH-M e IDEB de 2019.

correlação for negativa, ela indica que uma variável aumenta à medida que a outra diminui e vice-versa. A correlação também pode ser chamada de medida de associação, medida de interdependência, medida de intercorrelação ou medida de relação entre as variáveis (LIRA, 2004).

Existem três principais medidas de correlações: correlação de Pearson, correlação de Kendall, e correlação de Spearman. A correlação de Pearson é utilizada para verificar a relação entre variáveis lineares. A correlação de Kendal é um teste não paramétrico para medir a dependência entre duas variáveis. A correlação de Spearman é um teste também não paramétrico para medir o grau de associação entre duas variáveis.

A correlação também pode ser classificada como simples, múltipla ou parcial. A correlação simples analisa a dependência de duas variáveis, seja elas X e Y, sendo uma dependente e a outra independente. A correlação múltipla estuda a relação entre

uma variável dependente e outras duas ou mais variáveis independentes. A correlação parcial ocorre quando é realizada a correlação múltipla, após eliminar uma das variáveis independentes.

Nesta pesquisa, foi utilizada a análise de correlação simples de Pearson, por ser uma das mais comuns e mais usadas. Através da correlação de Pearson, obtém-se o coeficiente de correlação. O coeficiente fica em um intervalo de -1 a 1 e indica o grau de relação entre a variável que se obtém a correlação e a variável alvo, também chamada variável dependente (SOUSA, 2019). Se o valor da correlação for 0 (zero), ele indica que não existe associação entre as variáveis. Mas isso não significa que não existe uma relação entre elas.

No entanto, é importante destacar que, embora a correlação seja um método estatístico importante para verificar o relacionamento entre duas ou mais variáveis, a correlação não indica causalidade. Ou seja, se uma determinada variável tem uma correlação forte, conforme especificado na Tabela 6, não significa que tal variável seja a causa da variável com quem se relaciona. As variáveis podem ser influenciadas por um fator desconhecido.

Tabela 6 – Interpretações dos resultados de uma análise de correlação.

Correlação (+ ou -)	**Interpretação**
0.00 a 0.19	Correlação bem fraca
0.20 a 0.39	Correlação fraca
0,40 a 0.69	Correlação moderada
0.70 a 0.89	Correlação forte
0.90 a 1.00	Correlação muito forte

4.3.2 Resultados

A correlação foi realizada a partir de cada uma das variáveis com a variável dependente, qual é a nota do IDEB de 2019. Os dados foram divididos por microrregião e, em cada uma delas, buscou-se saber qual variável mais se correlacionava e a que menos se correlacionava com o IDEB a fim de comparar os resultados. As microrregiões são agrupamentos de municípios limítrofes, para integrar a organização, o planejamento e a execução de funções públicas de interesse comum (BRASIL, 1988).

Para a obtenção dos valores, foi utilizada a correlação de Pearson. No entanto, a correlação não implica causalidade, ou seja, forte correlação não é forte causa. Nas tabelas, as correlações mais altas estão destacadas em azul, enquanto as mais baixas em vermelho.

A Tabela 7 apresenta a correlação das microrregiões Litoral Ocidental Maranhense, Aglomeração Urbana de São Luís, Rosário e Lençóis Maranhenses. A partir desses resultados, destacamos os seguintes pontos:

- A microrregião Litoral Ocidental Maranhense teve como maior correlação a variável "Tem WI-FI", e como menor "Pais incentivam estudar";
- A microrregião Aglomeração Urbana de São Luís teve como maior correlação a variável "Escolaridade da Mãe" e como menor "Ler Quadrinhos";
- A microrregião Rosário teve como maior correlação "Escolaridade do Pai" e como menor "Ler Quadrinhos";
- A microrregião Lençóis Maranhenses teve como maior correlação a variável "Faz Dever de Casa" e como menor "Pais conversam sobre a escola".

Tabela 7 – Correlação das microrregiões Litoral Ocidental Maranhense, Aglomeração Urbana de São Luís, Rosário e Lençóis Maranhenses.

Variáveis	***Litoral Ocidental Maranhense***	***Aglomeração Urbana de São Luís***	***Rosário***	***Lençóis Maranhenses***
Escolaridade da Mãe	0.174347	0.650273	0.476202	0.250181
Escolaridade do Pai	0.205492	0.609390	0.524469	0.270106
Tem WI-FI	0.450420	0.564072	0.238235	-0.024634
Tem Computador	-0.128528	0.512691	0.199258	-0.035235
Pais conversam sobre a escola	0.171035	0.172491	0.345194	-0.055012
Pais incentivam tarefa de casa	0.183735	-0.193941	-0.150467	0.038601
Pais incentivam estudar	-0.343362	-0.142603	-0.009811	-0.040640
Pais incentivam ir à escola	0.040807	-0.129021	0.000397	0.244233
Pais comparecem a reuniões	0.011230	-0.055824	0.123554	0.308432
Ler Notícias	0.366475	0.054012	0.578248	0.377920
Ler Livros	0.360433	0.039617	0.386792	0.143884
Ler Quadrinhos	0.071692	-0.330629	-0.340454	0.075068
Evadiu	0.239202	0.490582	0.134037	0.442677
Reprovou	0.021073	0.422327	0.425448	0.226136
Idade que iniciou a estudar	-0.020444	0.318670	0.421277	0.353245
Faz Dever de Casa	-0.039998	0.175410	0.124839	0.489934

A Tabela 8 apresenta a correlação das microrregiões Gurupi, Codó, Coelho Neto, Caxias e Porto franco. A partir desses resultados, destacamos os seguintes pontos:

- A microrregião Gurupi teve como maior correlação a variável “Faz Dever de Casa” e como menor correlação a variável “Pais incentivam estudar”;
- A microrregião Codó teve como maior correlação “Pais comparecem a reuniões” e como menor “Ler Quadrinhos”;
- A microrregião Coelho Neto teve como maior correlação “Faz Dever de Casa“ e como menor “Pais incentivam estudar”;
- A microrregião Caxias teve como maior correlação “Tem WI-FI” e como menor “Pais incentivam tarefa de casa”;
- A microrregião Porto franco teve como maior correlação “Pais comparecem a reuniões” e como menor “Pais incentivam estudar”.

Tabela 8 – Correlação das microrregiões Gurupi, Codó, Coelho Neto, Caxias e Porto Franco.

Variável	***Gurupi***	***Codó***	***Coelho Neto***	***Caxias***	***Porto Franco***
Escolaridade da Mãe	0.477554	0.475201	0.250181	0.174347	0.049274
Escolaridade do Pai	0.188104	0.363611	0.270106	0.205492	0.221511
Tem WI-FI	0.248636	0.177342	-0.02463	0.450420	0.154569
Tem Computador	0.381700	0.229501	-0.03523	-0.12852	0.123266
Pais conversam sobre a escola	0.128439	0.263288	-0.05501	0.171035	0.058246
Pais incentivam tarefa de casa	0.006244	0.050850	0.038601	-0.34336	-0.212303
Pais incentivam estudar	-0.17339	0.191235	-0.04064	0.183735	-0.32333
Pais incentivam ir à escola	0.010561	0.298611	0.244233	0.040807	-0.223331
Pais comparecem a reuniões	-0.023769	0.476931	0.308432	0.011230	0.321884
Ler Notícias	0.157158	0.287085	0.377920	0.366475	0.129443
Ler Livros	0.474043	0.045738	0.143884	0.360433	0.308742
Ler Quadrinhos	-0.123178	-0.28549	0.075068	0.071692	-0.076696
Evadiu	0.192003	0.226103	0.442677	0.239202	0.248747
Reprovou	0.056870	0.374197	0.226136	0.021073	0.166379
Idade que iniciou a estudar	-0.007920	0.199380	0.353245	-0.02044	-0.204465
Faz Dever de Casa	0.517918	0.432436	0.489934	-0.03999	0.069851

A Tabela 9 apresenta a correlação das microrregiões Baixada Maranhense, Itapecuru Mirim, Pindaré e Imperatriz. A partir desses resultados, destacamos os seguintes pontos:

- A microrregião Baixada Maranhense teve como maior correlação "Escolaridade do Pai"e como menor "Ler Quadrinhos";
- A microrregião Itapecuru Mirim teve como maior correlação a variável “Ler Notícias” e como menor a variável “Ler Quadrinhos”;
- A microrregião Pindaré teve como maior correlação “Pais comparecem a reuniões” e como menor “Ler Quadrinhos”;
- A microrregião Imperatriz teve como maior correlação “Escolaridade da Mãe” e como menor “Ler Quadrinhos”.

Tabela 9 – Correlação das microrregiões Baixada Maranhense, Itapecuru Mirim, Pindaré e Imperatriz.

Variável	***Baixada Maranhense***	***Itapecuru Mirim***	***Pindaré***	***Imperatriz***
Escolaridade da Mãe	0.440658	0.385279	0.455201	0.611895
Escolaridade do Pai	0.613859	0.229740	0.393611	0.573775
Tem WI-FI	0.178394	0.446818	0.177342	0.427771
Tem Computador	0.550828	-0.159526	0.229501	0.564928
Pais conversam sobre a escola	0.131626	0.062308	0.253288	-0.062440
Pais incentivam tarefa de casa	-0.073118	0.337665	0.050850	-0.105393
Pais incentivam estudar	0.204034	-0.145638	0.181235	-0.053427
Pais incentivam ir à escola	0.292195	0.136207	0.288611	-0.092514
Pais comparecem a reuniões	0.261477	0.444723	0.456931	-0.087940
Ler Notícias	0.445384	0.693742	0.257085	0.265008
Ler Livros	-0.031569	-0.194759	0.045738	0.099506
Ler Quadrinhos	-0.202772	-0.507944	-0.285497	-0.260454
Evadiu	0.467724	0.351862	0.226103	0.156031
Reprovou	0.483920	-0.049711	0.364197	0.269640
Idade que iniciou a estudar	-0.034474	0.186183	0.199380	-0.022710
Faz Dever de Casa	0.177176	-0.239934	0.432436	0.456675

A Tabela 10 apresenta a correlação das microrregiões Médio Mearim, Alto Mearim e Grajaú, Presidente Dutra e Baixo Parnaíba Maranhense. A partir desses resultados, destacamos os seguintes pontos:

- A microrregião Médio Mearim teve como maior correlação a variável “Escolaridade da Mãe” e como menor “Ler Quadrinhos”;
- A microrregião Alto Mearim e Grajaú teve como maior correlação a variável “Pais comparecem a reuniões” e como menor “Idade que iniciou a estudar”;
- A microrregião Presidente Dutra teve como maior correlação a variável “Escolaridade da Mãe” e como menor a variável “Pais incentivam tarefa de casa”;
- A microrregião Baixo Parnaíba Maranhense teve como maior correlação a variável “Escolaridade da Mãe’ e como menor “Ler Quadrinhos”.

Tabela 10 – Correlação das microrregiões Médio Mearim, Alto Mearim e Grajaú, Presidente Dutra e Baixo Parnaíba Maranhense.

Variável	***Médio Mearim***	***Alto Mearim e Grajaú***	***Presidente Dutra***	***Baixo Parnaíba Maranhense***
Escolaridade da Mãe	0.560360	0.049274	0.562369	0.611895
Escolaridade do Pai	0.364618	0.221511	0.213996	0.573775
Tem WI-FI	0.261368	0.154569	0.392363	0.427771
Tem Computador	0.515942	0.123266	0.140927	0.564928
Pais conversam sobre a escola	0.022287	0.058246	-0.110567	-0.062440
Pais incentivam estudar	-0.225251	-0.136709	-0.108234	-0.053427
Pais incentivam tarefa de casa	-0.206242	-0.012303	-0.395472	-0.105393
Pais incentivam ir à escola	0.004400	-0.023331	-0.014667	-0.092514
Pais comparecem a reuniões	0.385724	0.321884	0.251748	-0.087940
Ler Notícias	-0.011099	0.129443	0.391236	0.265008
Ler Livros	-0.157589	0.308742	0.169343	0.099506
Ler Quadrinhos	-0.445228	-0.076696	0.029534	-0.260454
Evadiu	-0.005021	0.248747	-0.081090	0.156031
Reprovou	0.042253	0.166379	0.227293	0.269640
Idade que iniciou a estudar	0.072244	-0.204465	0.222161	-0.022710
Faz Dever de Casa	0.059788	0.069851	0.083014	0.456675

A Tabela 11 apresenta a correlação das microrregiões Chapadinha, Chapada do Alto Itapecuru, Gerais de Balsas e Chapada das Mangabeiras. A partir desses resultados, destacamos os seguintes pontos:

- A microrregião Chapadinha teve como maior correlação a variável “Escolaridade do Pai” e como menor “Ler Quadrinhos”;
- A microrregião Chapadas do Alto Itapecuru teve como maior correlação a variável “Ler Notícias e como menor “Ler Quadrinhos”;
- A microrregião Gerais de Balsas teve como maior correlação “Escolaridade da Mãe” e como menor “Pais incentivam ir à escola”;
- Chapada das Mangabeiras teve como maior correlação a variável “Escolaridade da Mãe” e como menor a variável “Pais incentivam ir à escola”.

Tabela 11 – Correlação das microrregiões Chapadinha, Chapada do Alto Itapecuru, Gerais de Balsas e Chapada das Mangabeiras.

Variável	***Chapadinha***	***Chapadas do Alto Itapecuru***	***Gerais de Balsas***	***Chapada das Mangabeiras***
Escolaridade da Mãe	0.490658	0.395279	0.495201	0.621895
Escolaridade do Pai	0.603859	0.259740	0.383611	0.583775
Tem WI-FI	0.178394	0.446818	0.177342	0.427771
Tem Computador	0.550828	-0.159526	0.229501	0.564928
Pais conversam sobre a escola	0.131626	0.062308	0.253288	-0.062440
Pais incentivam tarefa de casa	-0.073118	0.337665	0.050850	-0.105393
Pais incentivam estudar	0.204034	-0.145638	0.181235	-0.053427
Pais incentivam ir à escola	0.292195	0.136207	0.288611	-0.292514
Pais comparecem a reuniões	0.261477	0.444723	0.456931	-0.087940
Ler Notícias	0.445384	0.693742	0.257085	0.265008
Ler Livros	-0.031569	-0.194759	0.045738	0.099506
Ler Quadrinhos	-0.232772	-0.517934	-0.295497	0.260454
Evadiu	0.467724	0.361862	0.226103	0.156031
Reprovou	0.483920	-0.049711	0.364197	0.269640
Idade que iniciou a estudar	-0.134474	0.186183	0.199380	-0.022710
Faz Dever de Casa	0.187176	-0.239934	0.432436	0.456675

4.4 Regressão Linear da Relação do IDEB com o IDH-M

4.4.1 Conceitos

A regressão linear é uma tentativa de modelar uma equação que descreva o relacionamento entre duas variáveis (CURRAL, 1994). Um dos objetivos da regressão é realizar identificação e avaliação da relação entre uma variável dependente e uma ou mais variáveis independentes, também chamadas de preditoras ou explicativas. A regressão também pode ser aplicada para predizer valores futuros de uma variável (RODRIGUES; MEDEIROS; GOMES, 2013).

A regressão linear pode ser de dois tipos: simples ou múltipla. Simples é quando há uma única variável dependente e uma única variável independente. A regressão múltipla ocorre quando há mais de uma variável independente, e uma única variável dependente. Nesta pesquisa, foi usada a regressão linear simples.

4.4.2 Resultados

A Figura 7, a partir do gráfico de dispersão *scatter plot*, ilustra a regressão linear simples realizada com os dados do IDH-M (variável independente X) e os dados do IDEB 2019 (variável dependente Y). A reta traçada em vermelho representa os possíveis valores preditos de Y em função do valor já conhecido de X. Para medir a qualidade da regressão, foi utilizado o Erro Médio Quadrático (RMSE) que, ao ser calculado, utiliza a mesma escala da variável dependente (de 0 a 10). Nessa regressão, o RMSE foi de 0,43 e significa a diferença entre os valores preditos e os valores reais. O resultado ideal seria se o RMSE fosse 0 (zero), ou o mais próximo possível. Como visto, é possível verificar uma função linear com baixo RMSE que modela a relação entre as duas variáveis. Dessa forma, é possível predizer, com um erro aceitável, o valor do IDEB, em função do IDH-M de uma determinada cidade do estado do Maranhão.

4.5 Análise de Fatores

4.5.1 Conceitos

A Análise de Fatores é uma técnica estatística utilizada para reduzir a dimensão de um conjunto de dados em fatores em comum, visando diminuir o tamanho do conjunto de dados (GORSUCH, 2014). Através da análise de fatores, é possível tornar a dimensão de um determinado conjunto de dados bem menor. As variáveis com características em comum podem ser representadas por um único fator.

Os primeiros passos da análise de fatores foram dados por Charles Spearman e Karl Pearson, no entanto, também iniciada por outros pesquisadores renomados. A

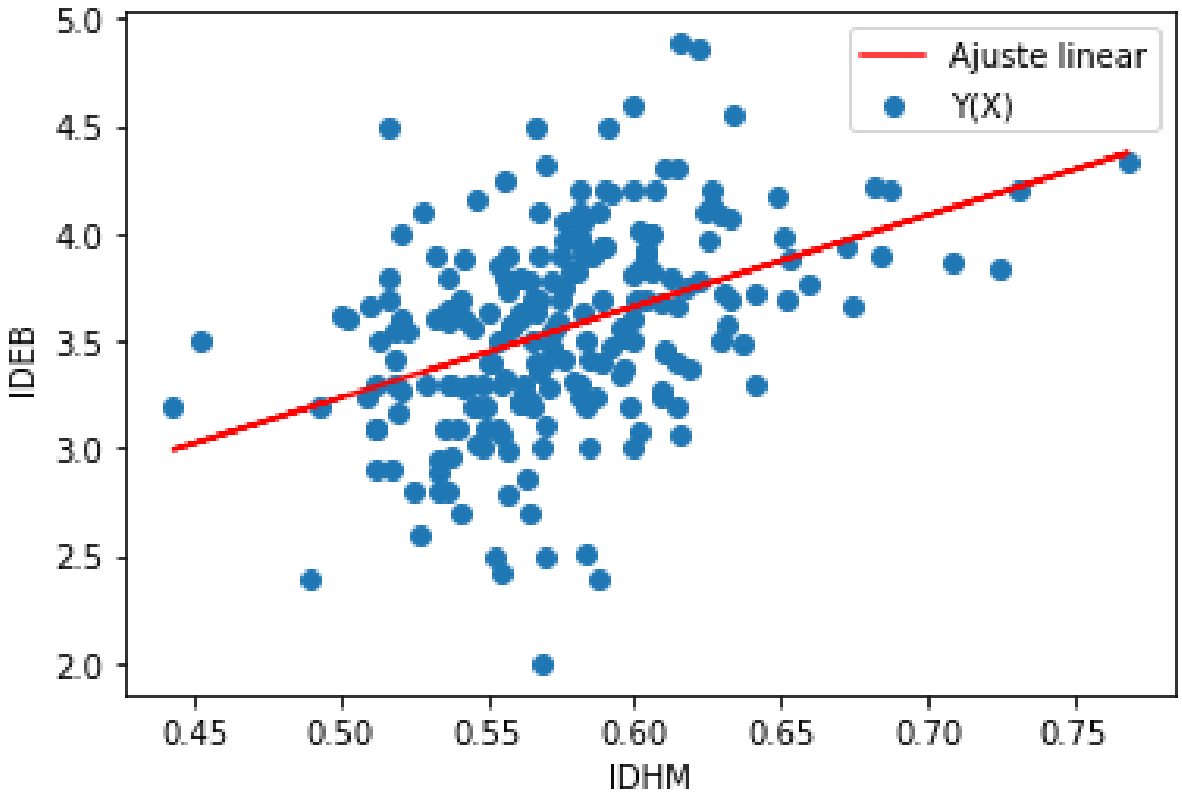

Figura 7 – Regressão linear considerando as variáveis IDH-M e o IDEB de 2019.

análise fatorial é uma técnica utilizada para, quando houver muitas variáveis observadas, gerar fatores subjacentes não observados. Dessa forma, o objetivo principal da análise fatorial é reduzir a quantidade de variáveis observadas. Através da técnica de análise de fatores, as variáveis podem ser condensadas em fatores comuns, tornando o conjunto de dados com menos variáveis, possibilitando sua análise e entendimento de forma mais fácil (DISTEFANO; ZHU; MINDRILA, 2009).

A análise fatorial pode ser de dois tipos: exploratória e confirmatória (FILHO; JÚNIOR, 2010). A Análise Fatorial Exploratória (AFE) é geralmente utilizada nos estágios iniciais da pesquisa para explorar os dados. Nessa fase, procura-se explorar a relação entre um conjunto de variáveis, identificando padrões de correlação. A Análise Fatorial Confirmatória (AFC) é utilizada para testar hipóteses. O pesquisador testa em que medida determinadas variáveis são representativas de uma dimensão. Nesta pesquisa, foi utilizada a análise fatorial confirmatória.

Para ocorrer a análise fatorial, o processo passa por um estágio de três fases (FILHO; JÚNIOR, 2010):

- Passo 1: Verificação de adequação da base de Dados: tamanho da amostra, variáveis contínuas ou discretas; exclusão de variáveis, como sexo, cor, por exemplo; limite *Kaiser-Meyer-Olkin* (KMO) mínimo de 0,6; e realização do Teste de Esfericidade de

Bartlett (do inglês, *Bartelett Test of Spherecity* - BTS) (p<0,05).

O teste KMO verifica se é apropriado usar as variáveis de manifesto para a análise fatorial. O teste realiza o cálculo da proporção de variância entre as variáveis de manifesto. Os valores de KMO variam entre 0-1, uma proporção abaixo de 0,6 sugere que o conjunto de dados é inapropriado para a análise fatorial. Já o Teste de Esfericidade de Bartlett é uma verificação de intercorrelação entre variáveis manifestas, ou seja, a comparação da matriz de correlação observada e a matriz de identidade. Se a análise fatorial for um método apropriado a ser usado, a matriz de correlação e a matriz de identidade não serão as mesmas e o teste será significativo (p<0,05).

- Passo 2: Determinar o número de fatores a serem extraídos através do *Scree test* (analisar graficamente a dispersão dos fatores) e do *Eigenvalue* igual ou acima de 1 (variância em todas as variáveis devida ao fator; variância explicada por cada fator da variância total; também é conhecido como raízes características).

Embora não exista uma regra absoluta de quantos fatores devem se extrair, a regra do *Eigenvalue* acima de 1 deve ser seguida, pois, se for abaixo desse valor, o fator pouco contribui para explicação das variáveis.

- 3. Decidir o tipo de rotação dos fatores: rotação Ortogonal do tipo *Varimax* ou rotação oblíqua.

A rotação é um método matemático que rotaciona os eixos no espaço geométrico com o propósito de facilitar a determinação de quais variáveis são carregadas em quais componentes. A rotação contribui para que o resultado empírico encontrado mais facilmente interpretável, conservando as suas propriedades estatísticas.

4.5.2 Resultados

A modelagem de análise de fatores foi realizada, pois, dentre as variáveis disponíveis, este estudo quer saber quais delas podem ser condensadas em fatores comuns que influenciam a nota do IDEB. Para fazer a análise de fatores, foi utilizado o software IBM SPSS. Após toda a preparação dos dados, o arquivo do tipo CSV foi importado para o software e executado os comandos de análise de fator. O software realiza todo o processo e também acrescenta ao conjunto de dados uma tabela com as pontuações dos fatores. O software IBM SPSS, em sua funcionalidade de análise de fatores, também apresenta a matriz de covariância, a matriz de correlações, Teste de KMO e Bartelett, Comunalidades, Variância total explicada, gráfico de escarpa, matriz de componentes, matriz de componentes rotativa, matriz de coeficientes de componente, gráfico de componente de fatores.

Através do gráfico de escarpa, também chamado *Scree Plot*, ou popularmente conhecido como gráfico de análise de cotovelo, pudemos avaliar a quantidade de fatores a serem resumidos a partir dos autovalores iguais ou superiores a 1. Essa mesma análise também pode ser feita através da matriz de variância total explicada onde mostra a porcentagem dos dados que podem ser explicados ou resumidos pelas principais variáveis mais relevantes.

Para efeito de elaboração dos fatores foram executadas e observadas as três fases necessárias, conforme descrito a seguir:

- A base de dados estava adequada, pois, ela foi importada totalmente pronta para o software com os dados devidamente tratados. Ao realizar o teste de KMO, o valor obtido foi 0,734, portanto um valor acima de 0,6, o que possibilitou fazer a análise de fatores;
- Ao realizar a análise do *Scree test*, foi evidenciado o total de 04 fatores;
- A rotação escolhida foi a Ortogonal *Varimax*, por ser a mais utilizada (PALLANT, 2020).

A Tabela 12 apresenta os fatores obtidos e agrupados em 04 grupos a partir das variáveis disponíveis na base de dados.

4.6 Árvore de Decisão

4.6.1 Conceitos

A árvore de decisão é um dos algoritmos supervisionados de aprendizado de máquina que considera a divisão dos dados em classes homogêneas para realizar uma classificação. O objetivo é encontrar atributos que geram a melhor divisão dos dados em subconjuntos com maior pureza, pertencente à classe alvo (GARCIA, 2003). Através da árvore de decisão, ilustra-se um mapeamento dos possíveis resultados de uma série de escolha, com a possibilidade de fazer comparações entre as opções propostas da árvore.

A árvore de decisão prioritariamente se origina a partir de um nó, que se subdivide em galhos com os possíveis resultados formando outros nós, e assim sucessivamente. Ela pode ser formada por três categorias de nós:

- Nó de decisão: mostra a decisão a ser tomada;
- Nó de probabilidade: mostra resultados incertos;
- Nó de desfecho: indica o resultado final.

Tabela 12 – Fatores identificados e variáveis correspondentes.

Fatores	***Variáveis***
Escolaridade dos Pais e Tecnologia	TEM COMPUTADOR
	TEM WI-FI
	ESCOL MAE
	ESCOL PAI
Incentivo dos Pais	PAIS CONVERSAM ESCOLA
	PAIS INCENTIVAM ESTUDAR
	PAIS TAREFA CASA
	INCENTIVAR IR ESCOLA
	PAIS COMPARECEM REUNIOES
Cultura	LER NOTICIAS
	LER LIVROS
	LER QUADRINHOS
Vida escolar	DEVER CASA
	IDADE INIC ESTUDAR
	REPROVOU
	EVADIU

Algumas vantagens da árvore de decisão podem ser destacadas, como: fácil compreensão; a árvore escolhe o melhor dentre as diversas opções de probabilidades; podem ser adicionadas novas opções e crescer a árvore de decisão; podem ser usadas combinadas com outras ferramentas que auxiliam na tomada de decisão. Dentre as desvantagens estão: os dados requerem uma preparação para serem usados na árvore de decisão, ela pode ser tornar bastante complexa e grande, dependendo da quantidade de dados.

4.6.2 Resultados

O algoritmo de árvore de decisão foi utilizado após ser feita a análise de fator. As variáveis foram condensadas em fatores, gerando um conjunto de dados com 04 variáveis correspondentes à pontuação de cada fator, e mais a variável alvo (notas do IDEB de 2019). Para fazer a árvore de decisão, foi utilizado software Orange Canvas.

Após a importação do conjunto de dados para o software, todas as variáveis foram categorizadas. Cada uma das variáveis correspondentes aos fatores foram categorizadas em quatro classes conforme sua influência baseada na pontuação de sua respectiva carga

de fator: muito baixo, baixo, bom e muito alto. A variável IDEB 2019 foi transformada em duas classes: abaixo da média e acima da média.

A Figura 8 apresenta o modelo de árvore de decisão resultante. Como pode ser visto, o fator "Escolaridade dos Pais", quando apresenta valores das classes bom e muito bom, tende à nota do IDEB ser acima da média; enquanto se esse fator for baixo ou muito baixo, a nota do IDEB será abaixo da média. O fator "Vida Escolar", quando possui valores nas classes bom e muito bom, tende a apresentar um IDEB acima da média; enquanto caso os valores estejam nas classes baixo ou muito baixo, o IDEB será abaixo da média.

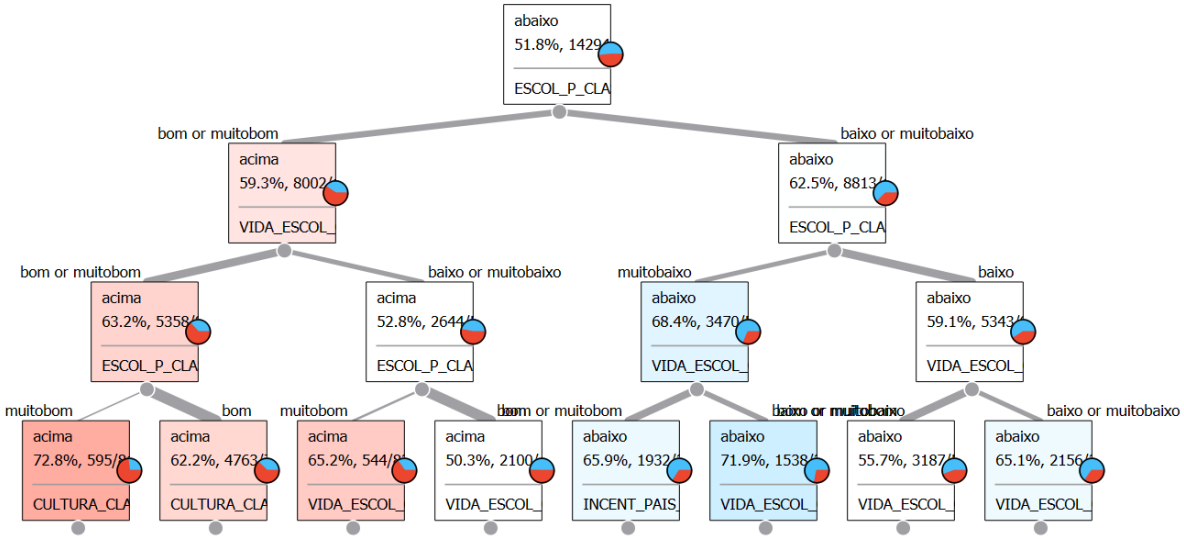

Figura 8 – Modelo de árvore de decisão gerado a partir dos fatores.

4.7 Conclusão

A etapa de preparação dos dados pode ser uma das mais demoradas, pois depende de como a base de dados se apresenta. Nesta etapa, os dados não estavam totalmente prontos, por isso, precisamos prepará-los para a mineração. Nesta pesquisa, a etapa de preparação dos dados foi realizada com uma complexidade que consideramos média. Após essa preparação, foi possível realizar a etapa de modelagem dos dados. No capítulo seguinte consta a avaliação das análises e modelos desenvolvidos.

5 Avaliação

Este capítulo informa se as questões de pesquisa foram respondidas, apresenta os principais achados da pesquisa e as dificuldades enfrentadas durante a sua execução. Também apresenta as limitações do estudo e os eventuais trabalhos futuros que poderão ser realizados.

5.1 Respondendo às Questões de Pesquisa e Verificação dos Critérios de Sucesso

A seguir são apresentadas as questões de pesquisas e as respostas para elas, e também a verificação de cumprimento dos critérios de sucesso, conforme o estudo realizado neste trabalho de mestrado.

- **(QP1)** Quais fatores mais influenciam o desempenho acadêmico das escolas de ensino médio do estado do Maranhão?
  Resposta: Escolaridade dos Pais e Tecnologia e Incentivo dos Pais. Os fatores encontrados foram descritos detalhadamente na seção 4.5.2.

- **(QP2)** É possível predizer a nota do IDEB a partir de um conjunto de fatores?
  Resposta: sim, a predição pode ser realizada através da realização de uma regressão linear ou árvore de decisão, analisando-se os resultados para verificar a acurácia da predição realizada.

Os **Critérios de Sucesso (CSs)** dessa pesquisa também foram alcançados, conforme descrito a seguir:

- (CS1) Houve o cumprimento dos objetivos almejados e a identificação das respostas para as questões de pesquisa;
- (CS2) Levantamento do estado da arte foi realizatravés da RSL;
- (CS3) Análise preliminar dos dados apresentados nas bases do SAEB e IDEB;
- (CS4) Identificação e análise dos fatores obtidos a partir da mineração dos dados;
- (CS5) A publicação de artigos científicos correspondente à etapa de distribuição da metodologia CRISP-DM é o único critério de sucesso ainda em fase de realização.

5.2 Principais Achados

Uma **análise preliminar** dos dados mostrou que a região metropolitana da capital São Luís é a que mais se destaca com os melhores indicadores educacionais. O IDEB das escolas da capital estão com notas que podem ser consideradas boas. No entanto, há escolas nas cidades do interior com IDEB acima da média. A região metropolitana da capital maranhense é melhor desenvolvida em alguns aspectos com disponibilização de uma infraestrutura que não está presente em outras regiões do estado. Embora não seja a melhor e nem a mais perfeita infraestrutura, isso acaba influenciando nos resultados dos indicadores educacionais. As regiões do interior do estado mostram um IDEB com índice baixo. O IDH-M desses municípios reflete também diretamente índices educacionais bons ou ruins. Quando o IDH-M da cidade é baixo, a tendência é que a nota do IDEB também seja abaixo da mediana, isso inclui a maioria das cidades do Maranhão. Quando o IDH-M é maior, o IDEB também tem uma tendência a ser acima da mediana, no entanto, esse índice se aplica a poucas cidades do Maranhão.

Ao **analisar a correlação** das variáveis selecionadas nesta pesquisa com as notas do IDEB, há uma grande variação de microrregião para microrregião. Não há como concluir sobre uma variável específica que tenha maior ou menor correlação. Tanto a variação do IDEB como a variação da correlação entre as variáveis, se comparados uma microrregião com a outra, ilustra a grande desigualdade e variações entre elas. Isso evidencia as disparidades regionais dentro do estado, pois não é possível identificar as mesmas variáveis como sendo de maior ou de menor correlação dentre as microrregiões. Portanto, cada microrregião tem suas particularidades. O Maranhão é um estado com extenso território, com peculiaridades diversas entre suas regiões.

No que diz respeito à **análise de fatores**, o principal fator que influencia na nota do IDEB é o fator da escolaridade dos pais, que considera o grau de estudo da mãe, pai ou responsável pelo aluno, e a tecnologia, que considera se o aluno tem computador e acesso à rede de Internet sem fio. Na análise de fatores, este foi o fator principal, destacando-se em primeiro lugar. O fator de incentivo dos Pais à educação dos filhos também é um fator bastante relevante para determinar os indicadores educacionais. Esse fator considera o nível de incentivo dos pais ou responsáveis no que diz respeito a participar de reuniões escolares, conversar com os filhos sobre a escola, incentivar a estudar, incentivar a fazer as tarefas de casa, e incentivar a ir à escola.

A **regressão linear** realizada com os dados do IDH-M e do IDEB 2019 mostrou-se capaz de realizar a predição das notas do IDEB. Mas deve se considerar que há uma variação muito grande entre as notas do IDEB de uma cidade para outra. A **árvore de decisão** realizada com as variáveis correspondentes à pontuação dos fatores em relação à variável alvo, notas do IDEB 2019, teve como nó raiz a escolaridade dos pais. Isso evidencia a escolaridade dos pais como um dos fatores determinantes e influentes na qualidade da

educação do Maranhão. Uma boa escolaridade dos pais ocasiona índices mais elevados do IDEB. Enquanto a escolaridade baixa dos pais pode implicar em baixo IDEB. Tando a regressão linear como a árvore de decisão podem ser usadas para predição de notas do IDEB, analisando-se os resultados para verificar a melhor predição ou a mais eficiente.

5.3 Dificuldades Enfrentadas

Esta seção apresenta as principais dificuldades encontradas no decorrer da pesquisa e escrita da dissertação. Uma das dificuldades enfrentadas nesse estudo foi no que diz respeito aos dados. Ao se trabalhar com dados de bases diferentes, tem-se que identificar uma variável comum aos dois (ou mais conjuntos de dados) para poder fazer a junção das bases de dados.

Outra dificuldade encontrada foi o fato das bases terem dados ausentes. Isso fez com que tenhamos que implementar alternativas para substituir esses dados, ou em muitos casos, excluir as linhas com esses dados por não ser possível aplicar nenhuma forma de substituição. Isso fez com que o conjunto de dados fosse reduzido.

A falta de padronização dos dados também atrapalhou, ou atrasar a pesquisa. Nesse sentido, os dados, para serem trabalhados, precisaram estar no mesmo padrão, requerendo tempo para ser feita essa padronização.

A realização de mineração de dados requer uso de tecnologias apropriadas. Algumas tecnologias estão disponíveis gratuitamente. No entanto, há tecnologias privadas, cuja disponibilização requer a compra de licença ou uso apenas durante o período de teste do software. No caso desta pesquisa o software privado utilizado foi o IBM SPSS.

5.4 Limitações do Estudo

Um estudo de mineração de dados é um processo iterativo e, consequentemente, ele vem com diversas limitações. Primeiramente, essa pesquisa se limitou apenas aos dados da rede pública estadual. A limitação aconteceu no sentido que, os dados do IDEB de escolas federais ou escolas particulares estavam ausentes. O motivo de dados ausentes se deve ao fato de que, as turmas dessas escolas, poucos alunos realizaram a prova do SAEB e, dessa forma, os dados não foram satisfatórios para cálculo da nota do IDEB. Em alguns caso, as escolas também solicitaram a não publicação dos seus dados do IDEB. Uma vez que a qualidade educacional pode variar (i.e., evoluir ou regredir) de um ano para outro (no caso do IDEB, com avaliação a cada dois anos), outra limitação foi a não realização de uma análise em série temporal, para verificar as mudanças que ocorreram ao longo dos últimos anos.

5.5 Trabalhos Futuros

Como trabalhos futuros, pode ser realizada uma comparação da qualidade educacional do Maranhão com a de outros estados e verificar quais fatores impactam a educação, por exemplo, dos estados do nordeste ou de outras regiões. Também pode ser feito um estudo comparativo dos dados do próprio Maranhão ao longo dos anos. Esta pesquisa de dissertação utilizou os dados do ano de 2019, estudos futuros podem usar os dados de novas edições ou de outras edições anteriores.

Como o Maranhão é um estado de território muito grande, haja vista a divisão em microrregiões, trabalhos futuros também podem ser realizados com dados regionalizados do Maranhão. A divisão pode ser em microrregiões ou mesorregiões, ou até mesmo de uma única cidade. Por fim, sugestões de trabalhos futuros podem ser realizadas mineração de dados para analisar o impacto da pandemia do Coronavírus (COVID-19) na educação maranhense, também através de uma análise comparativa de anos e inserção de outras bases de dados.

6 Considerações Finais

Este trabalho se propôs a identificar os fatores de influência da educação pública do Maranhão através de mineração de dados. Para isso, foi primeiramente realizado um estudo do estado da arte da mineração com os dados disponibilizados pelo INEP. Posteriormente, foram realizadas análises e desenvolvimento de modelos para estudar e entender a educação maranhense.

O objetivo do estudo foi alcançado e todos os critérios de sucesso foram atingidos, ou em processo de realização, incluindo, principalmente, a identificação das respostas para as questões de pesquisa. O estudo conseguiu identificar os fatores que influencia na educação básica do estado do Maranhão, utilizando-se de análise de fatores. Além disso, para uma melhor compreensão dos dados do IDEB, foram desenvolvidos alguns modelos que contribuíram diretamente ao cumprimento do objetivo e ao entendimento dos fatores de influência da educação do Maranhão.

Os resultados alcançados são relevantes no sentido que contribuem para um melhor entendimento da educação do Maranhão, considerando as notas do IDEB e também dos dados adquiridos através do questionário socioeconômico do SAEB. A pesquisa evidenciou que a vida socioeconômica do aluno influencia diretamente em sua vida escolar, podendo contribuir ou não para um bom desempenho escolar.

6.1 Contribuições Científicas

Considera-se que as principais contribuições desta pesquisa são:

1. A condução de uma revisão sistemática que serviu para mostrar o estado da arte em se tratando de mineração de dados com as bases de dados fornecidas pelo INEP para estudar problemas educacionais;

2. A identificação e listagem de fatores de influência na nota do IDEB que permite uma melhor compreensão de aspectos influenciadores nesse índice de qualidade educacional para o ano de 2019;

3. Essa pesquisa utiliza a metodologia CRISP-DM, cuja última etapa é a distribuição, em que deve ser mostrado o resultado da mineração de dados; os resultados desta pesquisa estão sendo distribuídos através da produção desta dissertação e também de um artigo científico para publicação.

6.2 Publicações

A partir do estudo realizado para esta dissertação de mestrado, foi possível realizar a elaboração de trabalhos científicos para publicação, a saber:

1. Inicialmente, foi realizado a publicação de um artigo com o conteúdo de uma RSL abordando a mineração de dados usando quatro bases de dados fornecidas pelo INEP (e.g, SAEB, IDEB, Indicadores Educacionais e Censo Escolar). Este artigo foi publicado na revista RENOTE;

2. Os resultados obtidos com o processo de mineração de dados educacionais realizado neste estudo serão apresentados em um artigo que atualmente está em produção a ser submetido para a RBIE;

3. Contribuição na elaboração de artigos em produção conjunta com outro aluno do Programa de Pós-graduação em Ciência da Computação (PPGCC); um artigo sobre a influência da COVID-19 nas notas do ENEM, que está submetido em processo de revisão por pares; uma RSL sobre aprendizado de máquina utilizando bases de dados educacionais, que está em fase de elabboração.

Referências

AGARWAL, S. Data mining: Data mining concepts and techniques. In: IEEE. *2013 International Conference on Machine Intelligence and Research Advancement*. [S.l.], 2013. p. 203–207. Citado na página 15.

AVILA, C. V. *et al.* Um linked data mashup de dados de execuções financeiras e indicadores educacionais no ensino básico. In: *Brazilian Symposium on Computers in Education*. [S.l.: s.n.], 2018. v. 29, p. 1911–1915. Citado 2 vezes nas páginas 26 e 31.

AZEVEDO, A. I. R. L.; SANTOS, M. F. Kdd, semma and crisp-dm: a parallel overview. *IADS-DM*, 2008. Citado na página 18.

BARBIERI, C. *BI–Business Intelligence–Modelagem & Tecnologia*. [S.l.]: Axel Books, 2001. Citado na página 26.

BEM, L. do N.; PEREIRA, V. da S.; SOUZA, E. Data mart para análise comparativa de dados do ideb em municípios da microrregião do pajeú em pernambuco. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. v. 6, p. 704–713. Citado 2 vezes nas páginas 26 e 31.

BIZER, C. *et al.* Ldif-a framework for large-scale linked data integration. 2012. Citado na página 26.

BRASIL. Constituição da república federativa do brasil de 1988. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Citado na página 47.

CARVALHO, J.; CRUZ, L.; GOUVEIA, R. Descoberta de conhecimento com aprendizado de máquina supervisionado em dados abertos dos censos da educaçao básica e superior. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. v. 6, p. 674–683. Citado 2 vezes nas páginas 25 e 31.

COHEN, J. *Statistical power analysis for the behavioral sciences*. [S.l.]: Routledge, 2013. Citado na página 24.

COLABORATORY, G. *Google Colaboratory*. 2020. <https://colab.research.google.com/notebooks/intro.ipynb>. Online; Accessed: Dec 12, 2020. Citado na página 39.

CURRAL, J. Statistics packages: A general overview. *Universidade de Glasgow*, p. 32, 1994. Citado na página 53.

DIMITOGLOU, G.; ADAMS, J. A.; JIM, C. M. Comparison of the c4. 5 and a naïve bayes classifier for the prediction of lung cancer survivability. *arXiv preprint arXiv:1206.1121*, 2012. Citado na página 26.

DISTEFANO, C.; ZHU, M.; MINDRILA, D. Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, v. 14, n. 1, p. 20, 2009. Citado na página 54.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996. Citado na página 15.

FILHO, D. B. F.; JÚNIOR, J. A. d. S. Visão além do alcance: uma introdução à análise fatorial. *Opinião pública*, SciELO Brasil, v. 16, n. 1, p. 160–185, 2010. Citado na página 54.

FONSECA, S. O. d.; NAMEN, A. A. Mineração em bases de dados do inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. *Educação em Revista*, SciELO Brasil, v. 32, p. 133–157, 2016. Citado na página 15.

GARCIA, S. C. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. 2003. Citado na página 56.

GONZALEZ, L. F. P. *et al.* Uma abordagem para mineração de dados e visualização de resultados em imagens batimetricas. Pontifícia Universidade Católica do Rio Grande do Sul, 2012. Citado na página 26.

GORSUCH, R. L. *Factor analysis: Classic edition*. [S.l.]: Routledge, 2014. Citado na página 53.

IBGE. *Maranhão - IBGE Cidades = <https://cidades.ibge.gov.br/brasil/ma/panorama>*. 2021. Online; Acessado em: 6 de maio de 2021. Citado na página 32.

IBM. *IBM SPSS Statistics*. 2022. <https://www.ibm.com/support/pages/ibm-spss-statistics-28-documentation>. Online; Accessed: Jan 25, 2022. Citado na página 39.

INEP. *Portal de Dados Abertos do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)*. 2020. <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos>. Online; Accessed: Dec 12, 2020. Citado 2 vezes nas páginas 35 e 36.

INEP. *Portal de Microdados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)*. 2020. <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados>. Online; Accessed: Dec 12, 2020. Citado 2 vezes nas páginas x e 33.

JÚNIOR, G. C. *et al.* Identificando correlaçoes e outliers entre bases de dados educacionais. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. v. 6, p. 694–703. Citado 2 vezes nas páginas 24 e 31.

JÚNIOR, O. d. G. F. *et al.* Melhoria da gestão escolar através do uso de técnicas de mineração de dados educacionais: um estudo de caso em escolas municipais de maceió. *RENOTE*, v. 17, n. 1, p. 296–305, 2019. Citado 4 vezes nas páginas 17, 28, 29 e 31.

JUNIOR, R. N. *et al.* Estimação de índices de aprovação e reprovação escolar do ensino médio. In: *Brazilian Symposium on Computers in Education*. [S.l.: s.n.], 2019. v. 30, p. 339–348. Citado 2 vezes nas páginas 28 e 31.

LIRA, S. A. Análise de correlação: Abordagem teórica e de construção dos coeficientes com aplicações. 2004. Citado na página 46.

MARANHãO, G. do. *Piso salarial do professor com jornada de 40 horas no Maranhão é R$ 3 mil a mais que o nacional*. 2022. <https://www.ma.gov.br/noticias/piso-salarial-do-professor-no-maranhao-e-r-3-mil-a-mais-que-o-nacional>. Online; Accessed: Apr 12, 2022. Citado na página 34.

MATPLOTLIB. *Matplotlib Documentation*. 2020. <https://numpy.org/doc/stable/>. Online; Accessed: Dec 12, 2020. Citado na página 38.

NAMEN, A. A.; BORGES, S. X. d. A.; SADALA, M. d. G. S. Indicadores de qualidade do ensino fundamental: o uso das tecnologias de mineração de dados e de visões multidimensionais para apoio à análise e definição de políticas públicas. *Revista Brasileira de Estudos Pedagógicos*, SciELO Brasil, v. 94, n. 238, p. 677–700, 2013. Citado na página 15.

NASCIMENTO, R.; JÚNIOR, G. C. Estudo sobre docentes do ensino básico através de indicadores educacionais e modelos de regressão. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2018. v. 7, p. 379–388. Citado 2 vezes nas páginas 27 e 31.

NASCIMENTO, R. L. do; FAGUNDES, R. A.; MACIEL, A. M. Prediction of school efficiency rates through ensemble regression application. In: *IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.: s.n.], 2019. v. 2161, p. 194–198. Citado 2 vezes nas páginas 27 e 31.

NASCIMENTO, R. L. S. do; JUNIOR, G. G. da C.; FAGUNDES, R. A. de A. Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. *RENOTE*, v. 16, n. 1, p. 1–11, 2018. Citado 2 vezes nas páginas 27 e 31.

NUMPY. *Numpy Documentation*. 2020. <https://numpy.org/doc/stable/>. Online; Accessed: Dec 12, 2020. Citado na página 38.

ORANGE. *Orange Documentation*. 2020. <https://orange.biolab.si/docs/>. Online; Accessed: Dec 12, 2020. Citado na página 39.

PACINI, I. B. de A. Educational indicators: a study of the limits and potentialities of the brazilian proof of the state teaching network of. *Humanidades & Inovação*, v. 7, n. 18, p. 242–257, 2020. Citado 2 vezes nas páginas 29 e 31.

PALLANT, J. *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. [S.l.]: Routledge, 2020. Citado na página 56.

PANDAS. *Pandas Documentation*. 2020. <https://pandas.pydata.org/docs/>. Online; Accessed: Dec 12, 2020. Citado na página 38.

PENTEADO, B. Geração automática de modelo de relações de pré-requisitos a partir de avaliação de larga escala brasileiras: um estudo preliminar. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2016. v. 5, p. 990–998. Citado 2 vezes nas páginas 23 e 31.

PENTEADO, B. E. Correlational analysis between school performance and municipal indicators in brazil supported by linked open data. In: *Proceedings of the 25th International Conference Companion on World Wide Web*. [S.l.: s.n.], 2016. p. 507–512. Citado 3 vezes nas páginas 17, 24 e 31.

PEñA-AYALA, A. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, v. 41, n. 4, Part 1, p. 1432–1462, 2014. Citado na página 15.

PINTO, G. da S.; JÚNIOR, O. d. G. F.; COSTA, E. de B. Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de teotônio vilela-alagoas. *RENOTE*, v. 17, n. 3, p. 183–193, 2019. Citado 2 vezes nas páginas 27 e 31.

PINTO, G. da S. *et al.* Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de maceió. In: *Brazilian Symposium on Computers in Education*. [S.l.: s.n.], 2019. v. 30, p. 1828–1837. Citado 2 vezes nas páginas 29 e 31.

PYTHON. *Python Documentation*. 2020. <https://www.python.org/doc/>. Online; Accessed: Dec 12, 2020. Citado na página 38.

RAMOS, T. G.; MACHADO, J. C. F.; CORDEIRO, B. P. V. Primary education evaluation in brazil using big data and cluster analysis. *Procedia Computer Science*, Elsevier, v. 55, p. 1031–1039, 2015. Citado 2 vezes nas páginas 23 e 31.

RIGOTTI, J. I. R.; CERQUEIRA, C. A. As bases de dados do inep e os indicadores educacionais: conceitos e aplicações. *Livros*, p. 71–88, 2015. Citado na página 15.

RODRIGUES, R. L.; MEDEIROS, F. P. D.; GOMES, A. S. Modelo de regressão linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2013. v. 24, n. 1, p. 607–616. Citado na página 53.

SANTOS, A.; MEDEIROS, F. P. A. de. Relationship of federal funding to ideb results in a state in brazil: an approach based on educational data mining. In: IEEE. *15th Iberian Conference on Information Systems and Technologies (CISTI)*. [S.l.], 2020. p. 1–4. Citado 2 vezes nas páginas 29 e 31.

SCIKIT-LEARN. *Scikit-Learn User Guide*. 2020. <https://scikit-learn.org/stable/user_guide.html>. Online; Accessed: Dec 12, 2020. Citado na página 39.

(SEDUC-MA), S. de Estado da Educação do M. *Maranhão mantém trajetória de crescimento e atinge 3,7 no Ideb, maior nota da história*. 2020. <https://www.educacao.ma.gov.br/maranhao-mantem-trajetoria-de-crescimento-e-atinge-37-no-ideb-maior-nota-da-historia/>. Online; Accessed: Oct 12, 2020. Citado na página 32.

SHEARER, C. The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000. Citado 2 vezes nas páginas ix e 18.

SILVA, W. J.; SOUZA, R. M.; CYSNEIROS, F. psda: A tool for extracting knowledge from symbolic data with an application in brazilian educational data. *Soft Computing*, Springer, v. 25, n. 3, p. 1803–1819, 2021. Citado 2 vezes nas páginas 29 e 31.

SOARES, J. F. Measuring cognitive achievement gaps and inequalities: The case of brazil. *International Journal of Educational Research*, Elsevier, v. 45, n. 3, p. 176–187, 2006. Citado 3 vezes nas páginas 17, 22 e 31.

SOARES, R. de C. *et al.* Mineracão de dados da educação básica brasileira usando as bases do inep: Uma revisão sistemática da literatura. *RENOTE*, v. 19, n. 1, p. 361–370, 2021. Citado 2 vezes nas páginas 17 e 21.

SOUSA, Á. Coeficiente de correlação de pearson e coeficiente de correlação de spearman: o que medem e em que situações devem ser utilizados? *Correio dos Açores*, Gráfica Açoreana, Lda, p. 19–19, 2019. Citado na página 47.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. [S.l.]: Ciência Moderna, 2009. Citado na página 15.

WALTENBERG, F. D.; VANDENBERGHE, V. What does it take to achieve equality of opportunity in education?: An empirical investigation based on brazilian data. *Economics of Education Review*, Elsevier, v. 26, n. 6, p. 709–723, 2007. Citado 2 vezes nas páginas 22 e 31.

WITTEN, I. H. *et al.* Practical machine learning tools and techniques. In: *DATA MINING*. [S.l.: s.n.], 2005. v. 2, p. 4. Citado na página 26.

Apêndices

APÊNDICE A – Artigo Publicado

Mineração de Dados da Educação Básica Brasileira Usando as Bases do INEP: Uma Revisão Sistemática da Literatura

Raimundo de Castro Soares, UFMA, soares.raimundo@ufma.br, 0000-0002-0904-0370
Nelson Weber Neto, UFMA, nelsonweberneto@gmail.com, 0000-0002-1136-768X
Luciano Reis Coutinho, UFMA, luciano.rc@ufma.br, 0000-0001-7996-7334
Francisco José da Silva e Silva, UFMA, fssilva@lsdi.ufma.br, 0000-0001-8339-3679
Davi Viana dos Santos, UFMA, davi.viana@lsdi.ufma.br, 0000-0003-0470-549X
Ariel Soares Teles, IFMA/UFMA/UFDPar, ariel.teles@ifma.edu.br, 0000-0002-0840-3870

Resumo: A análise de problemas educacionais brasileiros através da mineração de dados tem se tornado cada vez mais presente em estudos científicos. Neste contexto, esta Revisão Sistemática da Literatura (RSL) tem o objetivo de identificar os estudos focados em minerar os dados da educação básica brasileira produzidos pelo Instituto Nacional de Estudos Pedagógicos Anísio Teixeira (INEP). Os artigos foram obtidos por meio de buscas em fontes de artigos nacionais e internacionais, retornando um total de 410 estudos, dos quais 19 atenderam aos critérios de seleção. Adicionalmente, buscou-se entender como ocorre o processo de mineração desses dados. Os resultados mostram que os pesquisadores se interessam por esses dados para estudar os problemas da educação básica brasileira, sugerindo melhorias e tomadas de decisão para abordá-los.

Palavras-chave: Mineração de Dados, INEP, IDEP, SAEB, Censo Escolar, Indicadores Educacionais.

Data Mining of Brazilian Basic Education Using the INEP Bases: A Systematic Literature Review

Abstract: The analysis of Brazilian educational problems through data mining has become increasingly present in scientific studies. In this context, this Systematic Literature Review (SLR) aims to identify studies focused on mining the data of basic Brazilian education acquired by the Brazilian Institute of Studies and Educational Research Anísio Texeira (INEP). The articles were obtained by searching national and international article sources, so returning a total of 410 studies, of which 19 met the selection criteria. Additionally, we sought to understand how the data mining process occurs. The results shown that researchers are interested in such data to study the problems of Brazilian education, hence suggesting improvements and decision making to address them.

Keywords: Data Mining, INEP, IDEP, SAEB, School Census, Educational Indicators.

1. Introdução

O Instituto Nacional de Estudos Pedagógicos Anísio Teixeira*(INEP), um órgão do Ministério da Educação do Brasil, é responsável pela realização de avaliações e questionários para a coleta de dados educacionais relevantes. Estas coletas de dados são direcionadas ao corpo docente, corpo discente e gestores de todos os níveis de ensino. Ao serem analisados de forma minuciosa, esses dados podem mostrar o diagnóstico educacional sob vários aspectos, como a taxa de eficiência, infraestrutura escolar, formação dos professores, complexidade da gestão escolar, nível socioeconômicos dos alunos, dentre outros.

Todos os anos, o INEP realiza o Censo Escolar, sendo um importante instrumento de coleta de informações da educação básica e pesquisa estatística educacional brasileira.

*⟨https://www.gov.br/inep/pt-br⟩

DOI: https://doi.org/10.22456/1679-1916.118526

A educação básica compreende a educação infantil, ensino fundamental I e II, e ensino médio. Para a realização do Censo, o INEP conta com a colaboração das secretarias de educação municipais e estaduais, e devem participar escolas públicas e privadas de todas as etapas da educação básica e profissional. O Censo Escolar contribui para o INEP elaborar os Indicadores Educacionais.

O INEP também é responsável pela realização do Sistema de Avaliação da Educação Básica (SAEB) e do Índice de Desenvolvimento da Educação Básica (IDEB). As avaliações aplicadas pelo SAEB acontecem a cada dois anos. Provas de duas disciplinas são aplicadas, Língua Portuguesa e Matemática, para estudantes do 5º ano do ensino fundamental I, 9º ano do ensino fundamental II, e 3º ano do ensino médio.

O IDEB é um indicador que mede a qualidade da educação básica brasileira, calculado a partir da média da proficiência das provas de Língua Portuguesa e Matemática aplicadas pelo SAEB, padronizada entre 0 e 10, e o indicador de rendimento baseado na taxa de aprovação dos alunos. Além das avaliações, os questionários do SAEB também são aplicados aos alunos, professores e gestores da educação. Eles visam obter dados sobre a infraestrutura das escolas e aspectos administrativos.

Todos os dados produzidos por meio dessas mecanismos são disponibilizados pelo INEP através de uma página para acesso a dados abertos[†]. O volume de dados disponível é grande e tem sido analisado para gerar conhecimento para gestores, permitindo a realização de tomadas de decisão baseadas em evidências. Considerando isso, diversas pesquisas em mineração de dados educacionais têm sido conduzidas.

Devido aos trabalhos nesta área, é necessário agregar o conhecimento já publicado. A realização de Revisões Sistemáticas de Literatura (RSLs) pode auxiliar nesta agregação. A RSL realizada por (Coelho e Silveira, 2017) analisou a aplicação de *Deep Learning* na mineração de dados educacionais e análise da aprendizagem. (Santos; Ferreira e Miranda, 2017) conduziram uma RSL sobre dados educacionais abertos no Brasil. Já o mapeamento sistemático realizado por (Maschio *et al.*, 2018) abordou os estudos focados em mineração de dados educacionais do Brasil publicados em eventos nacionais. Diferente dessas revisões, a RSL reportada nesta pesquisa objetiva identificar os estudos focados em minerar os dados produzidos pelo INEP (i.e., os dados do SAEB, IDEB, Censo Escolar e Indicadores Educacionais), visando entender como este processo é realizado para estudar e sugerir melhorias aos problemas da educação básica brasileira.

O restante deste artigo está organizado como segue. A Seção 2 descreve a metodologia utilizada na condução desta RSL. A Seção 3 descreve os resultados da revisão, enquanto a Seção 4 os discute. Por fim, a Seção 5.

2. Metodologia

A produção dessa RSL foi baseada no protocolo desenvolvido por Kitchenham e Charters (Kitchenham e Charters, 2007). O guia estabelece três etapas de uma RSL: planejamento, condução e produção do relatório. Essas etapas foram suportadas por uma ferramenta online de condução colaborativa de RSLs, o *Parsif.al*[‡].

2.1. Questões de Pesquisa

Para alcançar o objetivo dessa RSL, as seguintes questões de pesquisa (QP) foram formuladas.

- (QP1) Quais dados do INEP (IDEB, SAEB, Censo Escolar ou Indicadores Educacionais) têm sido minerados?

[†] ⟨https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos⟩
[‡] ⟨https://parsif.al/⟩

DOI: https://doi.org/10.22456/1679-1916.118526

- (QP2) Quais problemas da educação básica brasileira estão sendo abordados?
- (QP3) Quais metodologias de mineração de dados estão sendo usadas nos estudos?
- (QP4) Quais as técnicas estão sendo empregadas na mineração de dados?

2.2. Estratégia de Busca

Duas strings de busca foram elaboradas e aplicadas, conforme a Tabela 1. Os termos utilizados nas strings foram escolhidos visando maximizar a quantidade de artigos sobre o assunto. No processo de seleção das fontes de artigos, foi dada prioridade para as bibliotecas digitais, revistas e eventos científicos com escopo da área de informática na educação. A pesquisa foi realizada no dia 17 de abril de 2021 em todas as fontes. Após os artigos terem sido obtidos, a etapa de remoção de duplicados foi feita com o suporte da ferramenta *Parsif.al*.

Tabela 1. Strings de busca utilizadas.

String de Busca	**Fontes de Artigos**
(“data analytics” OR “data mining” OR “data analysis”) AND (SAEB OR INEP OR IDEB OR“basic education” OR “school census” OR “educational indicators”)) AND (Brazil OR Brazilian)	ACM Digital LIbrary IEEE Digital Library ScienceDirect Web of Science
(dados OR data) AND (SAEB OR INEP OR IDEB OR “educação básica” OR “basic education” OR “censo escolar” OR “indicadores educacionais”)	Workshops do Congresso Brasileiro de Informática na Educação (CBIE) Revista Brasileira de Informática na Educação (RBIE) Revista Novas Tecnologias na Educação (RENOTE) Simpósio Brasileiro de Informática na Educação (SBIE)

2.3. Critérios de Seleção

Os critérios de seleção dos artigos são especificados na Tabela 2. De maneira independente, dois pesquisadores foram responsáveis por aplicar os critérios para selecionar os artigos. Inicialmente, eles realizaram a leitura dos títulos e resumos para selecionar os artigos. Ao final deste primeiro passo, o coeficiente de Kappa de Cohen (Cohen, 1968) foi calculado para identificar o nível de concordância entre os pesquisadores em relação aos resultados obtidos. Os artigos conflitantes foram então reavaliados de forma a buscar um consenso entre os pesquisadores. Quando necessário, um terceiro pesquisador contribuiu com discussões para sanar conflitos.

Tabela 2. Critérios de inclusão e exclusão de artigos.

Critérios de Inclusão (CI)	**Critérios de Exclusão (CE)**
(CI1): Artigos que realizam mineração dos dados do SAEB, IDEB, Censo Escolar e Indicadores Educacionais	(CE1): Artigos que realizam mineração de dados que não sejam do SAEB, IDEB, Censo Escolar e Indicadores Educacionais
(CI2): Artigos em inglês ou português	(CE2): Artigos em idioma diferente do inglês e português
(CI3): Artigos completos	(CE3): Literatura cinza (e.g., teses, dissertações, artigos curtos, capítulos de livro, relatórios técnicos)
	(CE4): Texto completo não disponível online

2.4. Avaliação de Qualidade

A qualidade dos artigos selecionados foi avaliada respondendo as cinco questões de qualidade (QQ) listadas abaixo. A pontuação dada a cada uma, conforme as respostas, foi a seguinte: sim = 1 ponto, parcialmente = 0,5 ponto, e não = pontuação zerada. A análise de qualidade foi realizada por dois pesquisadores.

- (QQ1): O estudo apresenta explicitamente o problema de pesquisa?

DOI: https://doi.org/10.22456/1679-1916.118526

- (QQ2): O estudo apresenta claramente o seu objetivo?
- (QQ3): Existe uma descrição adequada do contexto em que o estudo foi realizado?
- (QQ4): A mineração dos dados foi feita de maneira rigorosa?
- (QQ5): O estudo apresenta novos achados?

2.5. Extração de Dados

Com o objetivo de responder às questões de pesquisa, os dados apresentados abaixo foram extraídos dos artigos selecionados. Esta fase de extração de dados foi realizada por um pesquisador e verificada por outros pesquisadores com experiência na condução de RSLs, informática na educação e mineração de dados.

1. Título, ano e fonte do artigo.
2. Dados do INEP (i.e., se SAEB, IDEB, Censo Escolar ou Indicadores Educacionais relacionados a alunos, escola, professor e notas do IDEB).
3. Problema abordado.
4. Metodologia de mineração de dados.
5. Técnica empregada na mineração de dados.

3. Resultados

A Figura 1 ilustra o fluxo do processo de revisão. A pesquisa nas bases de dados resultou um total de 410 publicações, com apenas 3 artigos removidos por serem duplicados. Houve então a aplicação dos critérios de inclusão e exclusão aos 407 artigos. Após a seleção realizada independentemente por cada pesquisador, o coeficiente de Kappa de Cohen foi calculado, buscando verificar o grau de concordância. O resultado do Kappa foi de aproximadamente 0,5, o que representa uma concordância moderada (Viera e Garrett, 2005). Portanto, os artigos foram reavaliados pelos pesquisadores envolvidos na fase de seleção até chegarem a um consenso. Ao final, um total de 19 artigos foram selecionados.

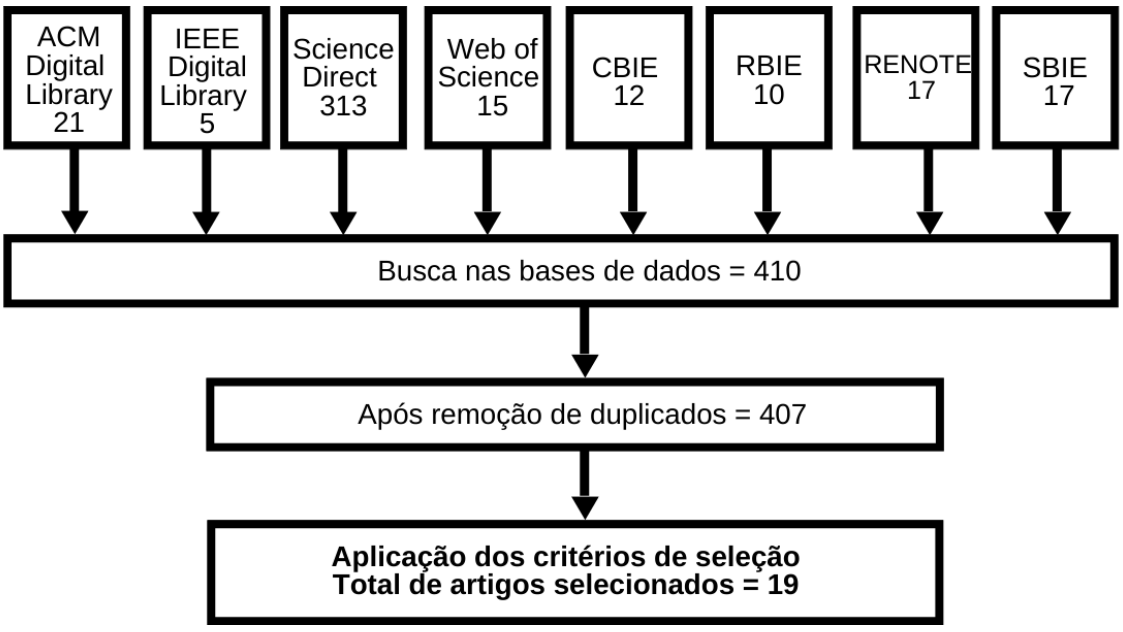

Figura 1. Fluxo do processo de condução desta revisão.

A Tabela 3 apresenta, em ordem cronológica de publicação, uma caracterização

DOI: https://doi.org/10.22456/1679-1916.118526

dos 19 estudos selecionados nesta RSL, com as seguintes informações: referências, dados analisados (i.e., SAEB, IDEB, Censo Escolar ou Indicadores Educacionais), os problemas estudados, as metodologias e as técnicas utilizadas.

Tabela 3. Caracterização dos estudos selecionados.

Referência	**Dados**	**Problema**	**Metodologia**	**Técnica**
A1 (Soares, 2006)	SAEB	Baixo desempenho escolar	Análise exploratória	Análise estatística
A2 (Waltenberg e Vandenberghe, 2007)	SAEB	Baixo desempenho escolar	Análise exploratória	Algoritmo EOp e Modelo de Regressão
A3 (Ramos; Machado e Cordeiro, 2015)	IDEB e SAEB	Fragilidade do sistema educacional	Análise exploratória	Análise de agrupamento
A4 (Penteado, 2016a)	SAEB	Melhoria de desempenho de alunos	Análise exploratória	Algoritmo POKS
A5 (Penteado, 2016b)	IDEB	Baixo desempenho escolar	Análise exploratória	Análise de correlação
A6 (Júnior *et al.*, 2017)	Censo	Baixo desempenho escolar	KDD	Análise de correlação e identificação de *outliers*
A7 (Carvalho; Cruz e Gouveia, 2017)	Censo	Evasão de alunos	KDD	Árvore de decisão e Classificação Bayesiana
A8 (Bem; Pereira e Souza, 2017)	IDEB	Falta de padronização entre bases educacionais	Proposta de solução	Ferramentas OLAP
A9 (Avila *et al.*, 2018)	Indicadores	Falta de padronização entre bases educacionais	Framework LDIF	Linked data mashup
A10 (Nascimento e Júnior, 2018)	Indicadores	Formação docente de baixa qualidade	CRISP-DM	Modelos de regressão
A11 (Nascimento; Junior e Fagundes, 2018)	Indicadores	Evasão e reprovação de alunos	CRISP-DM	Análise de correlação e Modelos de regressão
A12 (Nascimento; Fagundes e Maciel, 2019)	Indicadores	Aproximação do desempenho previsto das taxas de eficiência	CRISP-DM	Modelos de Regressão
A13 (Pinto; Júnior e Costa, 2019)	SAEB	Baixo Desempenho escolar	CRISP-DM	Algoritmos de classificação
A14 (Junior *et al.*, 2019)	Indicadores	Reprovação de alunos	Análise preditiva	Modelos de regressão
A15 (Júnior *et al.*, 2019)	IDEB e SAEB	Gestão de baixa qualidade e baixo desempenho escolar	CRISP-DM	Modelos de regressão e Árvore de decisão
A16 (Pinto *et al.*, 2019)	SAEB	Baixo desempenho escolar	CRISP-DM	Algoritmos de classificação
A17 (Pacini, 2020)	SAEB	Baixo desempenho escolar	Análise exploratória	Análise de correlação
A18 (Santos e Medeiros, 2020)	IDEB	Baixo investimento na educação	KDD	Análise de correlação e Modelo de regressão
A19 (Silva; Souza e Cysneiros, 2021)	SAEB e Censo	Qualidade do processo de mineração de dados	Proposta de solução	Modelo de regressão

Para o preenchimento da Tabela 3, algumas informações extraídas dos artigos tiveram que ser categorizadas e padronizadas. Ao relatar sobre os problemas educacionais abordados pelos estudos, o baixo desempenho escolar é uma categorização usadas para os estudos que não somente buscaram explicações para o baixo rendimento das escolas nos resultados do SAEB e IDEB, mas também sugerir melhorias.

As metodologias utilizadas pelos artigos foram: *Cross Industry Standard Process for Data Mining* (CRISP-DM), *Knowledge Discovery in Databases* (KDD), e a metodologia *Linked Data Integration Framework* (LDIF), uma ferramenta para a construção de aplicaç ões que utilizam dados relacionados (LDIF, 2021). Além disso,

DOI: https://doi.org/10.22456/1679-1916.118526

alguns estudos não definiram explicitamente uma denominação para a metodologia utilizada. Para esses estudos, categorizamos as metodologias de três formas: análise exploratória, usada para consultar a estrutura implícita dos dados e aprender sobre os relacionamentos entre as diversas variáveis; análise preditiva, cujo objetivo é analisar dados e classificá-los ou prever comportamentos futuros (e.g., notas futuras do IDEB); e dois artigos não especificaram exatamente uma metodologia, mas propuseram soluções para serem utilizadas na mineração de dados educacionais.

Na coluna que lista as técnicas empregadas na mineração dos dados educacionais, as ferramentas *Online Analytical Processing* (OLAP) se referem a um conjunto de técnicas e aplicações direcionadas ao acesso e análise *ad-hoc* de dados, tendo como objetivo transformar dados em informações para auxiliar no processo de decisão.

A Figura 2 apresenta o resultado da avaliação de qualidade dos artigos selecionados. Como pode ser visto, apenas 5 artigos obtiveram a pontuação máxima, e a pontuação mínima foi 3,5, atingida por 3 estudos. Esses resultados mostram que os estudos que utilizam os dados do INEP, em sua grande maioria, têm boa qualidade por, principalmente, cumprir seus objetivos.

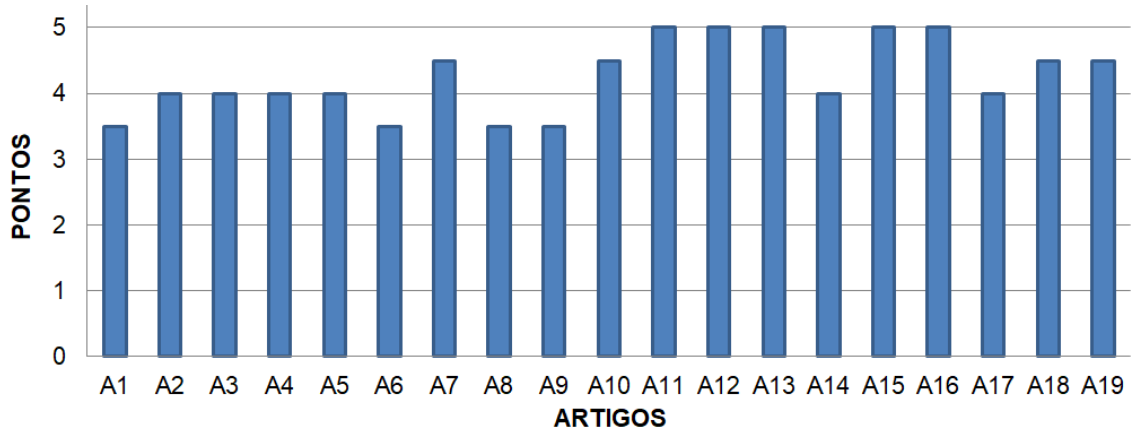

Figura 2. Pontuações dos estudos obtidas na avaliação de qualidade.

4. Discussão

4.1. Respondendo as Questões de Pesquisa

A questão de pesquisa **QP1** indagou quais os dados do INEP (IDEB, SAEB, Censo Escolar e Indicadores Educacionais) são analisados com técnicas de mineração. Constatou-se que, do total de artigos analisados, nove utilizam os dados do SAEB, sendo portanto o conjunto de dados mais utilizado. Cinco estudos utilizam dados dos Indicadores Educacionais, cinco utilizam os dados do IDEB e os dados menos utilizados são os do Censo Escolar, com apenas 3 artigos. Ressalta-se que, se somarmos a quantidade de vezes que os dados são utilizados, será maior que o total de 19 estudos selecionados, pois 3 desses trabalhos utilizam mais de um conjunto de dados do INEP (ver Tabelade resultados preliminares 3).

A questão **QP2** teve como objetivo identificar quais os problemas da educação básica brasileira estão sendo abordados pelas técnicas de mineração de dados. O principal problema estudado é o baixo desempenho escolar, o qual refere-se, por exemplo, ao baixo desempenho dos estudantes no SAEB e IDEB ou ENEM (Exame Nacional do Ensino Médio). O INEP fornece todos os anos uma estimativa sobre as notas do IDEB de acordo com os dados do ano anterior, desejando que as escolas alcancem aquelas notas para

DOI: https://doi.org/10.22456/1679-1916.118526

melhorar o nível da educação básica brasileira. É tácito que algumas escolas não alcançam essa estimativa. Dessa forma, o baixo desempenho escolar foi um problema frequente tratado nos artigos revisados. Também foram encontrado diversos outros problemas como a fragilidade do sistema educacional, evasão e reprovação de alunos, problemas com a formação dos docentes, baixo investimento na educação e dois artigos trataram da falta de padronização das bases de dados educacionais.

Quanto à **QP3**, a metodologia categorizada como análise exploratória de dados foi a mais utilizada. Quanto aos estudos que descrevem minunciosamente a metodologia utilizada, seguindo rigorosamente suas etapas, a mais utilizada foi a CRISP-DM, com 6 estudos deixando bem claro em seus textos que a utilizaram. Portanto, no que diz respeito a seguir rigorosamente as etapas definidas por uma metodologia, ela é a predominante e preferida para a realização de mineração de dados. A outra metodologia mais usada, que possui etapas definidas, foi a KDD, com 3 artigos utilizando-a. As metodologias CRISP-DM e KDD são bem parecidas no que diz respeito às etapas, então acreditamos que a escolha de uso pode ser devido às preferências dos pesquisadores.

Quanto à **QP4**, dentre as técnicas utilizadas, houve uma predominância de modelos de regressão e análise de correlação. Através de análise de coeficientes de correlação se define o grau de um relacionamento entre variáveis. Esta técnica é muito utilizada, por exemplo, na análise de dados de desempenho e infraestrutura escolar. Por outro lado, o desenvolvimento de modelos de regressão está também muito presente em diversos artigos. Esses modelos permitem aos pesquisadores fazerem uma predição do resultado (uma variável numérica dependente) e comparar com o resultado real.

A Figura 3 mostra a coocorrência das técnicas de mineração de dados do INEP utilizadas nos estudos revisados nesta RSL. Ela evidencia a resposta para a QP1, com os dados do SAEB sendo utilizados em maior número (n = 9) e a resposta da QP4, a qual identificou que as técnicas mais utilizadas são os modelos de regressão (n = 10) e análises de correlação (n = 5). Com a figura, é possível identificar ainda que os estudos têm focado mais frequentemente na elaboração de modelos de regressão com os dados dos Indicadores Escolares (n = 4) e do SAEB (n = 3).

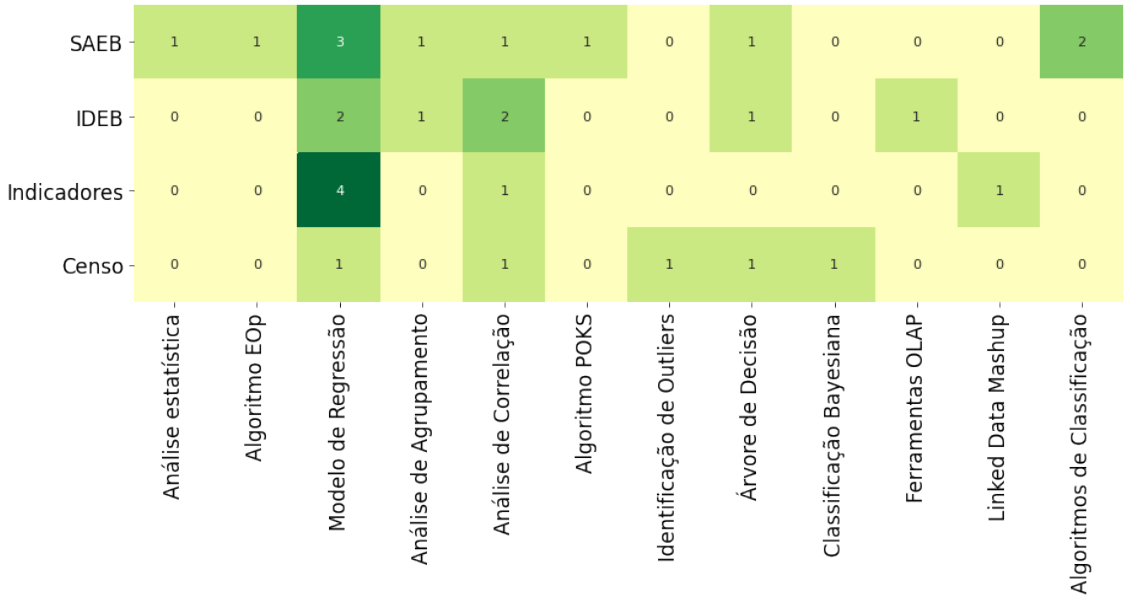

Figura 3. Coocorrência entre os dados do INEP (linhas) e as técnicas (colunas).

DOI: https://doi.org/10.22456/1679-1916.118526

4.2. Tendências

Existe uma tendência em grande parte dos artigos selecionados para tratar de problemas relacionados ao desempenho escolar de alunos. Minerando as bases de dados abertos do INEP, esses trabalhos visam identificar os fatores que influenciam no desempenho, criar modelos de predição, bem como sugerir possíveis soluções que podem ser feitas para melhorar as taxas de desempenho.

Outra tendência identificada foi a realização de análises exploratórias dos dados, sem um método bem definido nos artigos, e o uso da metodologia CRISP-DM. Quando utilizado aprendizado de máquina nos estudos, os modelos supervisionados de regressão foram os mais utilizados.

A Figura 4 faz uma análise entre o número de publicações por ano e os respectivos métodos de mineração de dados utilizados. Em relação ao quantitativo de publicações, fica evidente que no ano de 2019 houve um avanço significativo nas publicações. Em particular, muitos pesquisadores fizeram uso da metodologia CRISP-DM nos anos de 2018 e 2019. A figura ainda demonstra que, mesmo com o avanço nas tecnologias de mineração de dados, o ano de 2020 obteve um número baixo de publicações.

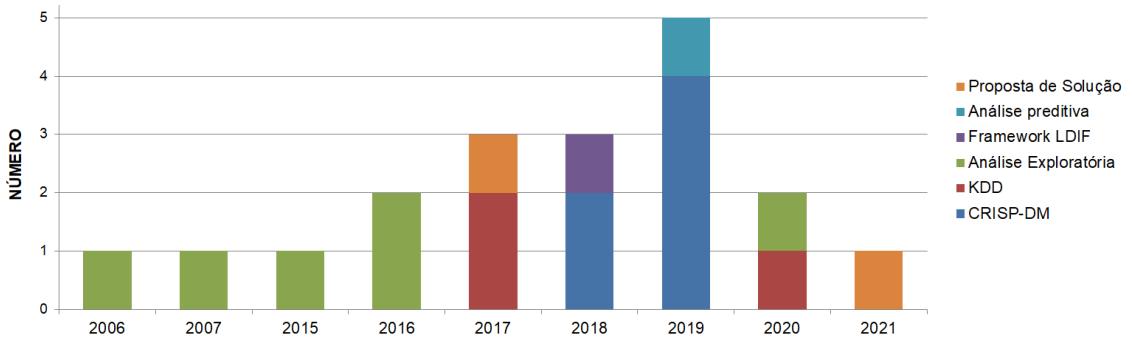

Figura 4. Quantidade de publicações por ano considerando as metodologias utilizadas.

4.3. Limitações

Algumas limitações relacionadas à condução desta RSL devem ser reconhecidas. Primeiramente, apenas dois idiomas de escrita dos artigos foram considerados. Dessa forma, artigos publicados em idioma diferente (e.g., espanhol) poderiam ser considerados relevantes mas não foram inclusos na revisão. Além disso, esta RSL tentou maximizar os resultados na estratégia de busca adotada, mas trabalhos não publicados (i.e., literatura cinza) que podem ter obtido resultados interessantes não foram considerados. Este trabalho se limitou à quatro bases de dados relacionadas à educação básica, porém o INEP disponibiliza outras bases, tais como o ENADE, ENEM, Censo dos Profissionais do Magistério e ENCCEJA, dentre outras. Portanto, percebe-se que ainda há oportunidade para a expansão desta RSL para outras bases do INEP e outros níveis da educação.

5. Conclusão

Este artigo teve como objetivo realizar uma RSL sobre mineração de dados da educação básica brasileira através da utilização dos dados do INEP. Um total de 19 artigos publicados foram analisados, com seus dados extraídos para responder às questões de pesquisa. Os dados do SAEB foram os mais utilizados nos estudos, mas todos os outros três conjuntos de dados (IDEB, Censo Escolar e Indicadores Educacionais) têm sido minerados pelas pesquisas inclusas nesta RSL. Diversos problemas educacionais foram identificados, tendo como predominância os problemas relacionados ao baixo

DOI: https://doi.org/10.22456/1679-1916.118526

desempenho escolar de alunos. Sobre as metodologias de mineração de dados utilizadas, não houve um método que predominou, mas a CRISP-DM foi a mais recorrente. Por fim, em relação às técnicas de mineração, pode-se concluir que, apesar de uma maior exploração do desenvolvimento de modelos de regressão, existe uma grande variedade das técnicas usadas.

Agradecimentos

Os autores gostariam de agradecer a Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) pelo apoio dado a seus projetos de pesquisa.

Referências

Avila, C. V. *et al.* Um linked data mashup de dados de execuções financeiras e indicadores educacionais no ensino básico. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2018. v. 29, p. 1911–1915.

Bem, L. do N.; Pereira, V. da S.; Souza, E. Data mart para análise comparativa de dados do ideb em municípios da microrregião do pajeú em pernambuco. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2017. v. 6, p. 704–713.

Carvalho, J.; Cruz, L.; Gouveia, R. Descoberta de conhecimento com aprendizado de máquina supervisionado em dados abertos dos censos da educaçao básica e superior. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2017. v. 6, p. 674–683.

Coelho, O. B.; Silveira, I. Deep learning applied to learning analytics and educational data mining: A systematic literature review. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2017. v. 28, p. 143–152.

Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. **Psychological bulletin**, 1968.

Júnior, G. C.; Nascimento, R.; Alves, G.; Gouveia, R. Identificando correlaçoes e outliers entre bases de dados educacionais. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2017. v. 6, p. 694–703.

Júnior, O. d. G. F.; Rodrigues, W. R. M.; Barbirato, J. C. C.; Costa, E. de B. Melhoria da gestão escolar através do uso de técnicas de mineração de dados educacionais: um estudo de caso em escolas municipais de maceió. **RENOTE**, v. 17, n. 1, p. 296–305, 2019.

Junior, R. N.; Nascimento, R. L. S. do; Fagundes, R. A. de A.; Neto, P. S. G. de M. Estimação de índices de aprovação e reprovação escolar do ensino médio. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2019. v. 30, p. 339–348.

Kitchenham, B.; Charters, S. **Guidelines for performing systematic literature reviews in software engineering**. [S.l.], 2007.

LDIF. **Linked Data Integration Framework**. 2021. ⟨http://ldif.wbsg.de/⟩. Online; Acessado em: 6 de maio de 2021.

Maschio, P.; Vieira, M. A.; Costa, N.; Melo, S. de; Júnior, C. P. Um panorama acerca da mineração de dados educacionais no brasil. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2018. v. 29, p. 1936–1940.

Nascimento, R.; Júnior, G. C. Estudo sobre docentes do ensino básico através de indicadores educacionais e modelos de regressão. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2018. v. 7, p. 379–388.

Nascimento, R. L. do; Fagundes, R. A.; Maciel, A. M. Prediction of school efficiency rates through ensemble regression application. In: **IEEE 19th International Conference on Advanced Learning Technologies (ICALT)**. [S.l.: s.n.], 2019. v. 2161, p. 194–198.

DOI: https://doi.org/10.22456/1679-1916.118526

Nascimento, R. L. S. do; Junior, G. G. da C.; Fagundes, R. A. de A. Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. **RENOTE**, v. 16, n. 1, p. 1–11, 2018.

Pacini, I. B. de A. Educational indicators: a study of the limits and potentialities of the brazilian proof of the state teaching network of. **Humanidades & Inovação**, v. 7, n. 18, p. 242–257, 2020.

Penteado, B. Geração automática de modelo de relações de pré-requisitos a partir de avaliação de larga escala brasileiras: um estudo preliminar. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2016. v. 5, p. 990–998.

Penteado, B. E. Correlational analysis between school performance and municipal indicators in brazil supported by linked open data. In: **Proceedings of the 25th International Conference Companion on World Wide Web**. [S.l.: s.n.], 2016. p. 507–512.

Pinto, G. da S.; Júnior, O. d. G. F.; Costa, E. de B. Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de teotônio vilela-alagoas. **RENOTE**, v. 17, n. 3, p. 183–193, 2019.

Pinto, G. da S.; Júnior, O. F.; Costa, E.; Barbirato, J. C. C.; Rodrigues, W. R. M. Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de maceió. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2019. v. 30, p. 1828–1837.

Ramos, T. G.; Machado, J. C. F.; Cordeiro, B. P. V. Primary education evaluation in brazil using big data and cluster analysis. **Procedia Computer Science**, Elsevier, v. 55, p. 1031–1039, 2015.

Santos, A.; Medeiros, F. P. A. de. Relationship of federal funding to ideb results in a state in brazil: an approach based on educational data mining. In: IEEE. **15th Iberian Conference on Information Systems and Technologies (CISTI)**. [S.l.], 2020. p. 1–4.

Santos, P.; Ferreira, R.; Miranda, P. Dados abertos educacionais: Uma revisao da literatura brasileira. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2017. v. 28, p. 11–20.

Silva, W. J.; Souza, R. M.; Cysneiros, F. psda: A tool for extracting knowledge from symbolic data with an application in brazilian educational data. **Soft Computing**, Springer, v. 25, n. 3, p. 1803–1819, 2021.

Soares, J. F. Measuring cognitive achievement gaps and inequalities: The case of brazil. **International Journal of Educational Research**, Elsevier, v. 45, n. 3, p. 176–187, 2006.

Viera, A. J.; Garrett, J. M. Understanding interobserver agreement: the kappa statistic. **Family Medicine**, v. 37 5, p. 360–3, 2005.

Waltenberg, F. D.; Vandenberghe, V. What does it take to achieve equality of opportunity in education?: An empirical investigation based on brazilian data. **Economics of Education Review**, Elsevier, v. 26, n. 6, p. 709–723, 2007.

DOI: https://doi.org/10.22456/1679-1916.118526